INCORPORATING MULTILEVEL GEOCODING AND SPATIAL MODELING TECHNIQUES TO PREDICT THE RISK OF WATER CONTAMINATION FOUND IN PRIVATE WELLS ACROSS GASTON COUNTY, NORTH CAROLINA

by

Claudio Owusu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Geography

Charlotte

2020

Approved by:

____

Dr. Eric Delmelle

____

Dr. David Vinson

____

Dr. Gary Silverman

____

Dr. Rajib Paul

Abstract

Claudio Owusu. INCORPORATING MULTILEVEL GEOCODING AND SPATIAL MODELING TECHNIQUES TO PREDICT THE RISK OF WATER CONTAMINATION FOUND IN PRIVATE WELLS ACROSS GASTON COUNTY, NORTH CAROLINA (Under the direction of Dr. Eric Delmelle)

Water is an important basic need for human survival. Many Americans obtain water from public water systems, however, about 45 million Americans use private wells for drinking water. When water is contaminated, it becomes unsafe for consumption and can cause many poor health outcomes. As a result, many environmental hazards present in water are regulated in public water systems, however, private well owners are not required to test, disinfect or treat their water.

In Gaston County, North Carolina, private wells provide the primary water supply for approximately 42% of the residents. Since 1989, well permits issued for new wells have been stored primarily on paper. Lack of digitization has hindered the ability of researchers and public health officials to access private well information. Further, lack of well testing data including arsenic and coliform has also made it difficult to determine groundwater quality in wells. No studies exist that describe spatial variation of arsenic and coliform bacteria presence in wells in Gaston County, that is within the NC Piedmont geologic belt.

The main objective of the dissertation is to incorporate multilevel geocoding and spatial modeling techniques to predict the risk of arsenic and coliform bacteria in private wells. To achieve this goal, first a GIS database of private wells is created using geocoding. Because the positional accuracy of private wells in GIS can affect the spatial analysis results, global positioning system (GPS) coordinates were obtained at 1035 wells to compare differences in the results of rooftop, parcel and street geocoding techniques. Second, the multilevel geocoding approach was used to determine the geographic coordinates of arsenic samples from 2011 to 2017. The sampled arsenic information, data on geology, pH, and well depth were used to estimate the probability of having arsenic at or above detectable levels ($\geq$ 5 µg/L) in wells across the county. This threshold was used because low levels of arsenic, even below the drinking water standard of 10 µg/L set by the United States Environmental Protection Agency, are still detrimental to human health and most of the arsenic detections in the study are between 5-10 µg/L. Third, coliform samples from private wells, well characteristics, parcel size, and soil ratings for the leachfield are examined to estimate the probability of having coliform bacteria in a well.

A multilevel geocoding approach improved match rate of permit addresses from 38.0% (n = 3,318) to 98.9% ($n = 8,616$). Addresses that were re-engineered during geocoding accounted for 50.9% ($n = 4,439$) of the matched records in the GIS database. There were significant differences ($p < 0.05$) in positional accuracy for rooftop, parcel, and street geocodes of private wells in the GIS database; positional accuracy was highest for rooftop geocodes.

Private wells set in mica schist (€Zms) were associated with arsenic at detectable levels, suggesting a local-scale geologic source influence of arsenic in the county. In addition, pH (median = 7.1) was positively associated with the presence of arsenic in well water, indicating arsenic $\geq$ 5 µg/L was predominantly associated with pH > 7.3. An area in the northwestern section (8.4 km^2) of the county was identified as having more than 50 percent likelihood of arsenic concentrations ≥ 5 μg/L. This area was found in the inner Piedmont of North Carolina belt and coincide with the mica schist geologic rock.

The multivariate logistic regression model results indicate that bored and older wells are more likely to have a high probability of coliform bacteria. The lack of significant association between poorly rated soils for a leachfield and probability of having coliform bacteria suggests that contamination is not a result of pathogens in household wastewater. There was no association between well depth and probability of having coliform in a well suggesting that contamination may come from runoff water.

Overall, the advanced geocoding approach can be used to improve geocoding match rate of input addresses for analytical purposes and develop a GIS database of private wells. The analysis of arsenic data in combination with geology, well depth and pH can provide preliminary insight into causes of long-term exposure to arsenic in groundwater. There was a higher chance of finding coliform bacteria in bored older wells. Because older wells (average well age = 19 years) were significantly likely to contain coliform bacteria suggest that those constructed before well standards was enforced may have a higher issue with coliform bacteria.

GIS maps can now be leveraged for targeted interventions to affected private wells in Gaston County. The present study is applicable to other regions interested in developing a GIS database of private wells, and towards advance understanding of spatial analysis of water hazards when few samples are taken in the field. This study provides a holistic approach that can be adopted for other regions facing similar groundwater exposure to environmental hazards in private wells.

ACKNOWLEDGMENTS

I never believed that I would be the first person in my family to attain a doctoral degree, considering I was not the smartest. However, with the demise of my brother, the honors fell on me to do whatever I can to live my dreams and, perhaps, also partially fulfill his. With a strong belief in God, perseverance, and support from my advisor, devoted committee members, family, and friends, I have finally completed my PhD.

I would like to express my sincere appreciation to my advisor Dr. Eric Delmelle for his profound belief in my abilities and his support throughout the journey to obtain this degree. I remember how we met in San Francisco, CA, after I lost out on a student paper competition, devasted but determined. I needed an opportunity to acquire a doctoral degree. Long story short, he invited me to a UNCC gathering, and we talked about a research opportunity called the "healthy wells project." Little did I know after gaining admission, I will work on this same project that has culminated in my dissertation. His supervision, practical suggestions, constructive criticism, patience, and friendly advice when needed have kept me going throughout these four years. I could not have imagined having a better advisor and mentor for my doctoral degree.

I am grateful to my committee members (Dr. David Vinson, Dr. Gary S. Silverman, and Dr. Rajib Paul) for all their support, critical comments, suggestions, and their enthusiasm to work with me on my dissertation. I am also very grateful to Dr. Wenwu Tang who, although not on my committee, was instrumental in my success at the Center for Geographic Information Science (CAGIS), where I worked tirelessly on my research. I am also grateful to Dr. Andy Bobyarchick for his review of the geologic data and collaboration. I would also like to acknowledge Drs. Douglass Shoemaker, Elizabeth Delmelle, Heather Smith, Deborah Thomas, Arif A. Ahmed, and Joseph Kangmennaang for all their support during my doctoral degree. I am also grateful to the entire faculty and staff of the Department of Geography and Earth Sciences for all their help whenever I needed it.

Thanks to Samantha Dye of Gaston County Health and Human Services for the collaboration on data and all the support throughout my time at UNCC. I am grateful to Dr. Kathleen M. Baker, Dr. Benjamin Ofori Amoah and Mrs. Agnes Ofori Amoah, for all the help you have given me since I have been in the United States.

To my very good friends Dr. Michael Richard Desjardins, Dr. Alex Hohl, Stephen Anim Preko, Ernest Otchere, Fred Owusu Agyeman, Delove Arthur, Daniel Twum, Remi Timbela, Eugene Keteni, Eugene Ahwireng, Eroy George Gyan, and Richard Atta Boateng am glad to have known you in my life and thanks for everything. I would also like to thank my CAGIS colleagues for their constant support and for creating an atmosphere of trust in which lab mates bounded with each other and built friendships.

My acknowledgment would be incomplete without thanking the prime source of my strength and happiness, which is my family. To my parents Mr. Kwesi Owusu Acheaw & Mrs. Juliana Addo, my sister, Rev. Rita Owusu Amponsah, and my brother, Bernard Owusu Amoako have all made marvelous contributions in my life and have helped me in many endeavors I cannot script down now. I also appreciate my close friends back in Ghana who consistently inspired me by their strides in life notably, Pius

Owusu Tufour, Dennis Awuah, Onyx Sonoh Adjei, Hans Tweneboah, and Amos Asamoah Boateng.

Most importantly, I wish to thank my phenomenal, loving and loyal wife, Rita Owusu (yes, my sister and wife have the same name) for coming to my life at a time when I needed support to complete my doctoral degree and start a family. And to our son, Micah Adams Owusu, thanks for choosing us to be your parents in this world. Finally, I would like to dedicate this work to my late brother Richy Owusu Kyere whose dreams for me have resulted in this achievement. I hope you rest well and see you someday in heaven.

April 2020

Charlotte, NC

DEDICATION

I dedicate this work to my late brother, Richy Owusu Kyere

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Environmental hazards

The World Health Organization (WHO) estimates that nearly one-quarter of all global deaths are due to environmental hazards (WHO 2016). Environmental hazards are chemical, biological and physical factors which have the potential to pose a threat to human health (Briggs, 2000). Examples of environmental hazards can include chemical factors such as arsenic in drinking water or soil, biological factors such as the presence of respiratory viruses in the air (e.g. Coronavirus), and physical factors related to access to clean drinking water, flooding and earthquakes.

Sometimes, the exposure to the environmental hazards may occur over a prolong period before the individual can experience any health effect (Briggs, 2000). For example, Li et al. 2013 found that individuals exposed to arsenic levels < 10 µg/L for more than 10 years through drinking water experience an increased risk of hypertension and type 2 diabetes. Similarly, long term exposure to radiation have been related to increase for thyroid cancer and other neoplasms (Schneider & Sarne, 2005). Also, acute exposure to environmental hazards can be dangerous to human health. For example, direct exposure to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the air can cause Coronavirus disease 2019 (COVID-19), which can lead to severe illness and death (Dyal, 2020; Heinzerling et al., 2020).

The conditions that put people at risk for environmental hazards vary over space and time, which translates into geographic variation in vulnerability to environmental hazards and health effects (Maantay and McLafferty 2011). Individual mobility and space-time variation of hazards create patterns of exposure and risk. Cromley and McLafferty (2011) reiterate that geography can provide valuable insights into how environmental factors influence health.

Health geography attempts to understand the interaction between people and their environment, which includes environmental health hazards (Dummer 2008). As a result, the relationship between environment and health is conceptualized into three distinct links: the environment, the population, and health (Briggs, 2000). To understand the interconnected links, it is important to understand that certain environmental factors create the hazards. When people encounter hazards and become exposed, their health deteriorate. And a population health effects can also be described when the exposed individuals live within close geographic areas. Health problems are then analyzed from a spatial perspective to understand how geographic context of a space and connection between places plays a major role in shaping environmental risks and health outcomes (Macintyre et al. 2002; Cutchin, 2007; Lu and Delmelle 2019).

Due to the emphasis on space and location, geographic information systems (GIS) and spatial methods are uniquely suited to environmental hazards research (Maantay and McLafferty 2011). For example, GIS has long been used in mapping the distribution and magnitude of environmental hazards (Cromley and McLafferty 2011). Also, GIS and spatial methods has been used in evaluating the potential risk to human health, either by estimating the numbers of people exposed to the environmental hazard or estimating the likely health burden (He et al., 2020) Further, GIS can be used in mapping the actual health outcome which can be attributed to exposures to the environmental hazards of interest (Meliker et al., 2010).

The analysis of environmental hazards that affect water quality is an important theme in health geography because of the many health benefits derived from drinking clean water. For example, water helps to get rid of waste in the form of urine from the body and is essential for cooking food. Arsenic and coliform are two common hazards regulated in public drinking water sources due their significant health impacts in both acute and chronic exposure in drinking water. For example, arsenic has been identified as a known carcinogen that causes several cancers (International Agency for Research on Cancer, 2004). On the other hand, coliform in drinking water may indicate that harmful pathogens are present, and these microorganisms may cause waterborne diseases including cholera, polio and typhoid fever (Centers for Disease Control and Prevention (CDC), 2016).

Many studies have found high levels of arsenic in groundwater sources compared to surface water sources of drinking water (Pippin 2005; Ayotte et al. 2006; Harden et al. 2009). This is due to groundwater interaction with arsenic sources including the geology. On the other hand, whereas coliform occur naturally in the environment in soils and plants and in humans and animals, they may be found in groundwater when wastewater leaks into the groundwater. The presence of coliform in groundwater may suggests that contamination sources are nearby the drinking water (Gerba 2009; Cabral 2010). This study is to understand the potential factors and spatial patterns of arsenic and coliform in an area with high usage of groundwater for drinking water.

1.2 Public Water Systems and Private Wells

Safe drinking water is essential for maintaining good health and reducing mortality from consuming untreated water (Centers for Disease Control and Prevention

(CDC) 2019). Access to clean water is related to many factors including where you live and socioeconomic status. For example, spatial analysis of the Flint water crises in 2014 showed that exposure to elevated blood lead levels was associated with living in low socioeconomic neighborhoods (Hanna-Attisha et al. 2016). Further, access to clean water may be impeded by having environmental hazard sources in proximity to water systems including public water systems and private wells. The later however, is prone to environmental hazards because the source comes from groundwater.

In order to protect the public from exposure to elevated levels of environmental hazards in drinking water in the United States (U.S.), the U.S. Environmental Protection Agency (U.S. EPA) under the Safe Drinking Water Act (SWDA) of 1974 must regulate public water systems (PWS) (U.S. EPA 2019a). To qualify as a PWS, drinking water must be supplied to at least 25 people. Under the SWDA, operators of PWS are required to schedule frequent testing of the quality of the water to ensure compliance with the maximum contaminant levels (MCL) set by the USEPA, but the regulations do not apply to private wells (Tiemann 2014).

Although most Americans drink water from PWS, about 13 million households (45 million Americans) rely on unregulated private wells for drinking water (USEPA 2019a). Compared to the rest of the U.S., some mid-west states (Michigan, Wisconsin, and Illinois), some northeast states (Pennsylvania, New Hampshire, Vermont, and Maine) and Montana have more than 25 percent of their population using private wells (*Figure 1.1*a). Other states such as Idaho, Wyoming, south Atlantic states (Maryland, West Virginia, South Carolina and North Carolina) as well as Minnesota, and Connecticut have from 21 to 25 percent of the population using private wells for drinking water wells (Figure 1.1a).

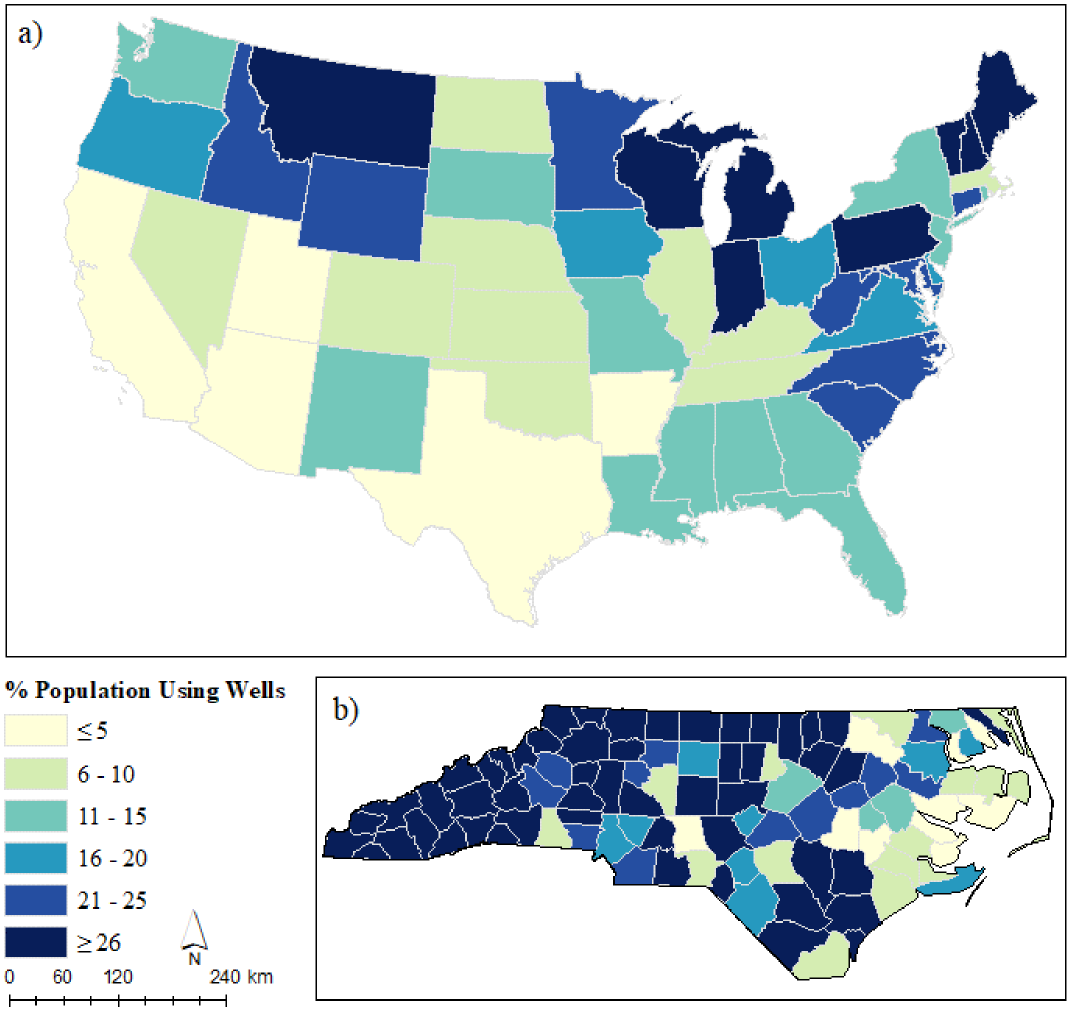

Figure 1.1:The percentage of people using private wells a) per state in conterminous United States, and b) per county in North Carolina (Data source: Dieter et al., 2018).

Compared to the rest of North Carolina, most counties in the western, northern and southern parts of North Carolina have more than one-fourth of their residents using private wells for drinking water (*Figure 1.1*b). Counties with ≤ 20 percent of population using private wells are in the east part of North Carolina (*Figure 1.1*b). Because private wells are unregulated, users are solely responsible for the safety of their drinking water.

In North Carolina, new private wells must be tested for bacteria and chemical contaminants at the time of construction (MacDonald Gibson and Pieper 2017; North Carolina Department of Health and Human Services 2019), and while further contaminant testing is encouraged, it is not required.

Several studies have reported that well owners infrequently test for contaminants (Knobeloch et al. 2013; Swistock et al. 2013; Pieper et al. 2015). Yet, private wells are at risk of exposure to environmental hazards found naturally in rocks or soils, septic systems, hazardous sites, landfills, pesticides and many other sources (USEPA 2015). Yet, private well owners are not required to monitor environmental hazards (Fox et al. 2016).

On a national scale, the United States Geological Survey (USGS) produces reports of groundwater quality every year; however, their analysis does not include data from private wells (Maupin et al. 2014). Instead, water quality from sentinel wells is used in the report. Sentinel wells are established to maximize the information provided, which is different from sampling household wells. For example, numerous sentinel wells may surround a hazardous waste site, instead of household residential area. This study is an advance towards having a comprehensive understanding of groundwater quality in an area with large usage of private wells.

1.3 GIS-based Surveillance Systems

Many existing inventories of private wells in the U.S. lack digital geographic coordinates, and county-level permitting system often store information in paper copies (MacDonald Gibson and Pieper 2017). This creates two main challenges; (1) inability to determine locations of private wells, and by extension (2) inability to characterize the spatial and space-time pattern of environmental contaminants. To overcome these challenges, linkages of electronic database of permits and laboratory chemical and biological test results from private wells are needed. This can be facilitated by using a Geographic Information System (GIS).

GIS-based surveillance of private wells can enable the storing, querying, and estimation levels of environmental hazards across an area (Cromley and McLafferty 2011). This is possible by integrating and analyzing data on sources of contaminants and locations of private wells in GIS. Because GIS data of private wells mostly represents individual residences, the population affected by the contamination can also be accurately estimated.

Several studies have utilized GIS techniques in exposure assessment (Christakos and Serre 2000; Reif et al. 2003; Sanders et al. 2012; Navoni et al. 2014). For example, GIS can be employed in the preliminary steps of delineating a target population in an exposure assessment process (Elliott et al. 2001; Cockings et al 2004; Weis et al. 2005). Similarly, a population's exposure to arsenic-contaminated water can be defined using geographic boundaries stored in GIS databases. GIS-based techniques can be used to specify both potentially exposed and control groups, especially in studies requiring individual analysis over an extensive focus area (Elliott et al. 2001; Nuckols et al. 2011). Further, web-based GIS tools have been used in volunteer data collection to reduce sampling bias and monitor lead (Pb) levels in Flint Michigan (Goovaerts, 2017a, 2017b; Abokifa, Katz, & Sela, 2020).

These examples further show that GIS can be useful in getting geographic coordinates, integrating data in different formats (e.g., satellite image, shapefile) from multiple sources, help with planning sampling, monitoring levels of contamination in environmental hazards assessments. The research conducted in this dissertation is to understand a systematic approach to generate geographic coordinates from paper records, integrate multiple datasets on sources and factors that account for the high likelihood of arsenic and coliform in wells. As a result, regular update of the data and results in this study is an advance towards the creation of a GIS-based surveillance system to monitor groundwater quality.

1.4 Arsenic Contamination

Arsenic contamination is a major global human health problem, and it is estimated that 140 million people are currently at risk of arsenic-related diseases as a result of drinking contaminated water (World Health Organization (WHO) 2019). In 2019, arsenic was ranked as the first substance of priority in the U.S. based on the known or suspected toxicity and potential for human exposure (Agency for Toxic Substances and Disease Registry (ATSDR) 2020). Among the 45 million Americans using private wells, 2.1 million individuals use water above the drinking water arsenic standards of 10 µg/L set by the USEPA (Ayotte et al. 2017).

Immediate ingestion of arsenic into the body causes some enzymes to be inactive, particularly those involved in cellular pathways, DNA synthesis, and immune formation (Ren et al. 2010; Flora and Medicine 2011). Lifetime chronic exposure to arsenic at elevated levels (>10 µg/L) in water has been related to several types of cancers including prostate (Benbrahim-Tallaa and Waalkes 2008), lung (Heck et al. 2009; Dauphiné et al.

2013), bladder (Steinmaus et al. 2003), kidney (Yuan et al. 2010), and skin (Karagas et al. 2015).

Recent studies have suggested that even low levels of arsenic (≤ 10 µg/L) in drinking water may impact fetal development (Bloom et al. 2016; Almberg et al. 2017), increase odds of diabetes (Li et al. 2013), and cause heart diseases (Bräuner et al. 2014; James et al. 2015). These studies have suggested that lowering the regulatory limit of arsenic may reduce the potential effects on human health when consumed at low levels. Unnoticed long-term and short-term exposure to arsenic in wells has the potential to lead to adverse health outcomes and is a public health concern (Ayotte et al. 2017).

The spatial distribution of arsenic in well water is often concentrated in distinct geographic regions or areas (Sanders et al. 2012). For instance, using data compiled from 31,000 groundwater samples by the USGS, it is clear that compared to the rest of the U.S., the western states have higher levels of arsenic concentrations greater than 10 µg/L (Figure 1.2a). Also, parts of the Midwest states have arsenic greater than 10 µg/L in the groundwater (Figure 1.2a). From the map (Figure 1.2a)., it appears that arsenic concentrations is lower in the southeast of the U.S., but this may be due to the small amount of samples collected from this area. For example, the map of North Carolina in *Figure 1.2*b appears to show that most areas have lower than 10 µg/L arsenic concentrations in groundwater.

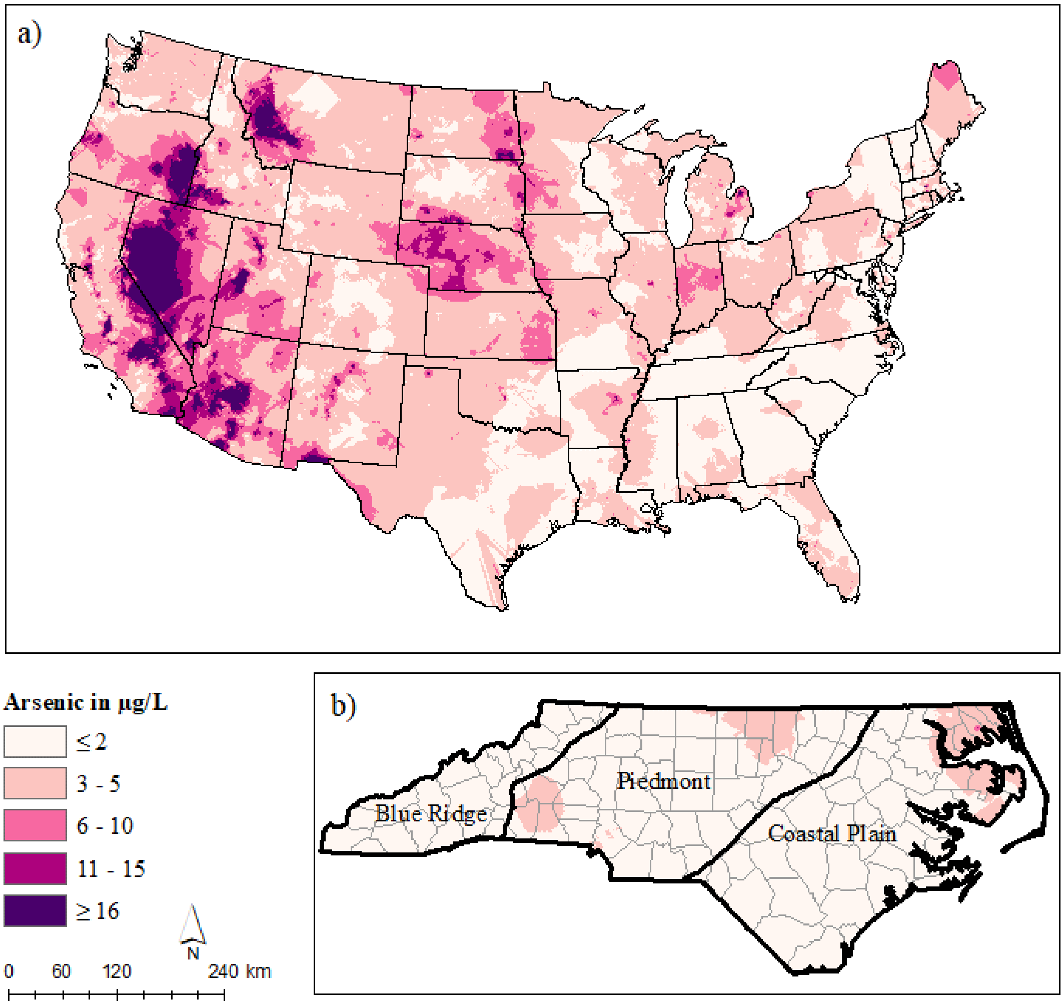

Figure 1.2: Arsenic in groundwater for, a) conterminous United States, and b) North Carolina showing the physiographic regions (Data source: Data.Gov, 2020)

Further research in North Carolina using detailed samples indicate that the Piedmont of North Carolina, which includes part of Gaston County, has elevated levels of arsenic (Pippin 2005; Reid et al. 2005). In aquifers of the Piedmont of North Carolina, elevated arsenic concentrations have been related to metavolcanic or metavolcaniclastic rocks (Pippin 2005; Harden et al. 2009). Chapman et al. (2013) have also found associations between elevated arsenic in well water and metamorphosed clastic sedimentary rocks. Arsenic dissolves out of the bedrocks when groundwater levels drop through evaporation (Centers for Disease Control and Prevention (CDC), 2015). Chapman et al. (2013) also found that elevated arsenic concentrations were positively associated with pH (measure of the acidity - low pH or alkalinity – high pH) in well water. When water is at a high pH, the formation of soluble ions can increase arsenic mobilization through desorption processes (Ayotte et al. 2003; 2006).

Well depth has also been identified as a significant predictor of elevated arsenic concentrations, particularly for wells within welded tuffs and quartz units and those close to transition zones (Kim et al. 2011). The studies above have generated evidence that geologic sources, pH, and well depth may influence the arsenic concentration in wells. Due to the complexity of geologic units within North Carolina, it is expected that geographic distribution of arsenic concentration varies at a local scale of a county. This study is to understand the factors that influence arsenic levels in wells and characterize areas at risk of having significant high likelihood of detecting arsenic ≥ 0.5 μ/L due to the health effects at chronic exposure even at this low level.

1.5 Coliform Bacteria

Coliform bacteria are microorganisms found in humans and animals feces and in the environment (Farrell-Poe et al. 2010). Although some coliform bacteria may not be harmful, generally their presence in drinking water can indicate contamination with human or animal waste, and harmful pathogens could also be present (Gerba 2009; Cabral 2010; Pandey et al. 2014). Pathogenic microorganisms can cause diseases if the water is not adequately treated (Wallender et al. 2014; Beer et al. 2015; Benedict et al. 2017). Illnesses that result from consuming pathogenic contaminated water includes acute gastrointestinal illness, acute respiratory illness, and neurologic illnesses (Benedict et al.

2017). Waterborne diseases may even lead to early death (Cortese and Parashar 2009; Morgan et al. 2015). For example, Chaudhry et al. (2015) found that the presence of hepatitis E in drinking water of pregnant women increased their risk of fulminant hepatitis, deaths and may be fatal to the unborn baby.

To reduce health risks associated with drinking pathogenic contaminated water, public water systems must regularly test for total coliform in the U.S. (USEPA 2017). However, no testing is required for private wells even when they are close to pollution from septic systems (Kaplan 2014; USEPA 2015), and landfills (Charrois 2010), raw manure (Sadeghi and Arnold 2002) and animal grazing (Hubbard et al. 2004; Fairbrother and Nadeau 2006).

Several studies have found that within communities with a high density of septic systems, areas with small parcels are more likely to test positive for coliform bacteria in well water compared to wells in areas with large parcels (Patterson 1999; Knierim et al. 2015). This is because putting many leachfields next to each other above the same aquifer in an area with small parcels can increase the risk of incomplete removal of contaminants and impurities from wastewater in the soil (Yates 1985; McQuillan 2004; Swartz et al. 2006).

Furthermore, the removal of contaminants and impurities from wastewater in leachfields have been found to be efficient in soils with good purification abilities (Beal et al. 2005). The USDA Natural Resources Conservation Service (2019) includes it in the computation of soil rating for a leachfield. The soil rating for a leachfield may serve as an essential variable to understand pathways of contamination in private wells but has not been yet examined in studies which attempt to predict the probability of having coliform bacteria in well water.

Previous studies have provided valuable insights that well age is positively associated with wellhead failures and may serve as a route of entry for surface contamination (Gonzales 2008; Sarkar et al. 2012). Moreover, coliform bacteria are more likely to occur in bored wells (Conboy and Goss 2000; Olabisi et al. 2008; Maran et al. 2016). Bored wells may have a higher likelihood of having coliform bacteria present in the water because they tap groundwater from the shallow unconsolidated material which is encountered by dirty runoff water when it rains (Mesner 2012). This dissertation attempts to capitalize on all the various factors to predict the probability that coliform bacteria is present in a private well.

1.6 Study Area

Gaston County, North Carolina (Figure 1.3B) has a landmass of 364 mi² or 942km² and is a fast-growing county of nearly 225,000 residents (2019) in the South-Central Piedmont section of North Carolina. The county is bounded on the east by the Catawba River and Mecklenburg County, on the west by Cleveland County, on the north by Lincoln County and on the south by York County, South Carolina. Gaston County enjoys a temperate climate with moderate temperature variations and humidity. The topography of the County is gently rolling to hilly, with several pronounced ridges. Elevations above sea level range from 587 feet (179 meters) in the southeast corner to 1,705 feet (520 meters).

The CDC recently reported that 42% of residents in the county use private wells (CDC 2019). From 1989, the Gaston County authorities required a well owner to duly be approved a permit before a private well is constructed (The Gaston County Board of Health 2011). Noteworthy is that private wells constructed prior to 1989 did not require permits as a result most of them may not have the same construction standards introduced in the Gaston County well ordinance (The Gaston County Board of Health 2011).

Private well permits issued from 1989 to date by the Gaston County Department of Health & Human Services (GC-DHHS) are kept as paper copies. Private wells information prior to 1989 are unknown because issue of a well construction permit was not required. Geocoding the addresses on paper permits is critical for the creation of a GIS database of private wells, but the paper nature of the permits presents many challenges to the digitization process. Some of the paper permits were particularly challenging to digitize and geocode due to damaged paper, missing or non-specific address information, directional addresses, and illegible handwriting.

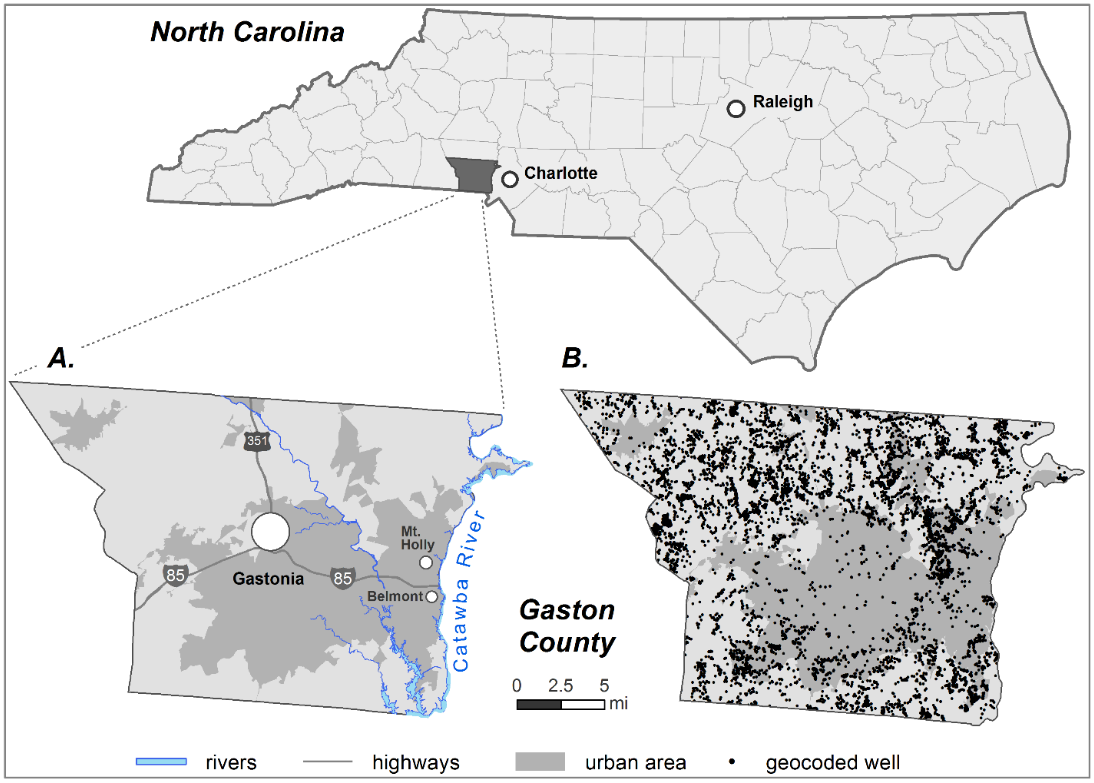

Figure 1.3: Map of the study area (A: Gaston County; B: spatial distribution of geocoded wells)

These types of challenges present difficulties in the development of GIS databases of private wells needed across the U.S. to monitor groundwater quality influenced by disparate hazard sources. This study uses a multistage-geocoding approach to determine the geographic coordinates of the wells from the paper permits. As shown in Figure 1.3B, are the geocoded locations of the private wells. Most of the private wells are found outside the urban areas suggesting that rural residents use private wells for drinking water compared to urban residents (Figure 1.3B). This is because most urban residents rely on public water systems in the county.

Gaston County is representative of the many rural counties in the US where a variety of pollution sources affect groundwater quality for millions of Americans that use private wells as drinking water source. The results of this dissertation are therefore portable to other areas in the country where development of a GIS database is foundational to subsequent geographic analysis of environmental hazards in private wells.

1.7 Limitations of Previous Work and Contributions of the Dissertation to GIScience and Health Geography

The application of geographic methods has long been a core component of environmental hazards research (Maantay and McLafferty 2011). For example, hazard sources are classified by the geography of discharge process using GIS (Cromley and McLafferty 2011). Using GIS, the discharge from a single location can be shown as a point, and nonpoint discharge sources as lines or area features in GIS (Nuckols et al. 2004). GIS can be used to locate the origin of water pollution such as arsenic and subsequently trace the various pathways of exposure towards determining individual exposure levels (Reif et al. 2003; Navoni et al. 2014). Because the reliability of geographic analysis depends on the quality of input data (Zandbergen 2009), there has been a consistent effort to improve GIS data.

Further, because studies vary in terms of context – for example, the scale of analysis can vary among studies, thus existing methods may need to be retested with new data. Nonetheless, new methods evolve over time to advance the field of GIScience and health geography. This dissertation identifies gaps in geographic studies in geocoding, uncertainty (focus on positional accuracy) and application of existing spatial techniques in environmental hazard assessment of groundwater quality in private wells. Further discussion on the research gaps and contributions of the dissertation in the geographic field are outlined in the sections below.

1.8 Geocoding

Most GIS-based studies geocoded data, global positioning system (GPS) collected data and, digitized data from public agencies and aerial images for spatial analysis (Cromley and McLafferty 2011). Due to cost and time, the most widely used approach to acquire spatial data is address geocoding (Goldberg et al., 2013). Address geocoding is the process of converting location information in the form of addresses into geographic coordinates of longitude and latitude (Goldberg et al. 2007). The method is available in most commercial GIS software's and as a result, the individual address is becoming a standard level of analysis for spatial investigations such as exposure assessment (Meliker et al. 2010), healthcare services (Delmelle et al. 2013) and identifying vulnerable areas of diseases (Owusu et al. 2018).

Because of the dependence on address geocoding for analytical purposes, decisions made based on such research may be impacted by having many ungeocoded records (Zimmerman, 2008). Ha et al. (2016) found that excluding ungeocoded records in the analysis reduces the sample size and may weaken the generalization of the analytical results due to selection bias. Many geographers have approached this problem from different viewpoints. Hart and Zandbergen (2013) have suggested the number of geocoded records can be improved by varying the spelling sensitivity of street names during geocoding. A drawback of this approach may be potentially selecting a wrong match address. Another widely used approach involves combining different reference datasets for geocoding the input addresses. Murray et al. (2011) combined street network and parcel data in geocoding to increase the match rate of the addresses of sex offenders in Ohio from 80 to 90 percent. This multi-stage geocoding approach involves arranging the reference data (i.e., rooftop centroids, parcel, and street networks) in a hierarchical order based on their spatial accuracy (Sonderman et al. 2012). Goldberg (2011) has also suggested that when only the street network data are available for geocoding, one can search for the probable address within a set of nearby candidates. This approach can be performed by utilizing online geocoding systems and through the manual placement of geocodes within GIS (McDonald et al. 2017).

Although these approaches have contributed to strategies that can improve the geocoding match rate, they fail to explore problems that could be inherent in the input addresses. For example, none of the studies proposed an approach to deal with missing or incomplete addresses. Data for geocoding often contain addresses and may also have additional attributes that can be linked to other reference datasets (e.g., parcels) to extract address information. This process can be facilitated by using a data-matching algorithm called probabilistic record linkage (PRL). PRL is used to match two datasets with similar attributes by assigning weights based on the degree of similarity (Randall et al. 2013). High weights suggest a higher probability of a match (Schmidlin et al. 2015). PRL has been used in health services research on birth outcomes and hospitalization records (Bentley et al. 2012). In this dissertation, the missing address on the paper permits were re-engineered using PRL. Because the utility of PRL could lead to improving input addresses, in this dissertation, the hypothesis is that it would translate into improving the geocoding match rate of private wells. Achieving a high match rate would mean that selection bias is reduced when results are used in further analysis, such as sending field teams to gather data on coliform bacteria in this dissertation. Reducing the gap in knowledge with regards to improving input addresses from public records is a critical contribution of this dissertation to GIScience.

1.8.1 Uncertainty

An emerging theme in the field of geography is the concept of uncertainty. Uncertainty is the difference between the digital representation of an object in the real world and the object itself (Zhang and Goodchild 2002). Goodchild (2009) identified sources of uncertainty to include errors, accuracy, and vagueness in definitions used in the compilation of geographic data. Many geographers have recognized that output from GIS based address geocoding has inherent error when compared to the actual position of homes and property boundaries on the ground (Bonner et al. 2003; Ward et al. 2005; Zimmerman et al. 2007; Zandbergen 2009; Jacquez 2012;). Error assessment in this situation is also referred to as positional accuracy and tends to differ by the reference dataset used in geocoding. The three main types of reference datasets used to convert an input address into longitude and latitude are rooftop points (i.e., rooftop centroid), parcels (i.e., parcel centroid), and street networks (Zandbergen 2008). As shown in Figure 1.4, the variation in the reference datasets, which translates to differences in the placement of the final geocoded output – result in differences in positional accuracy.

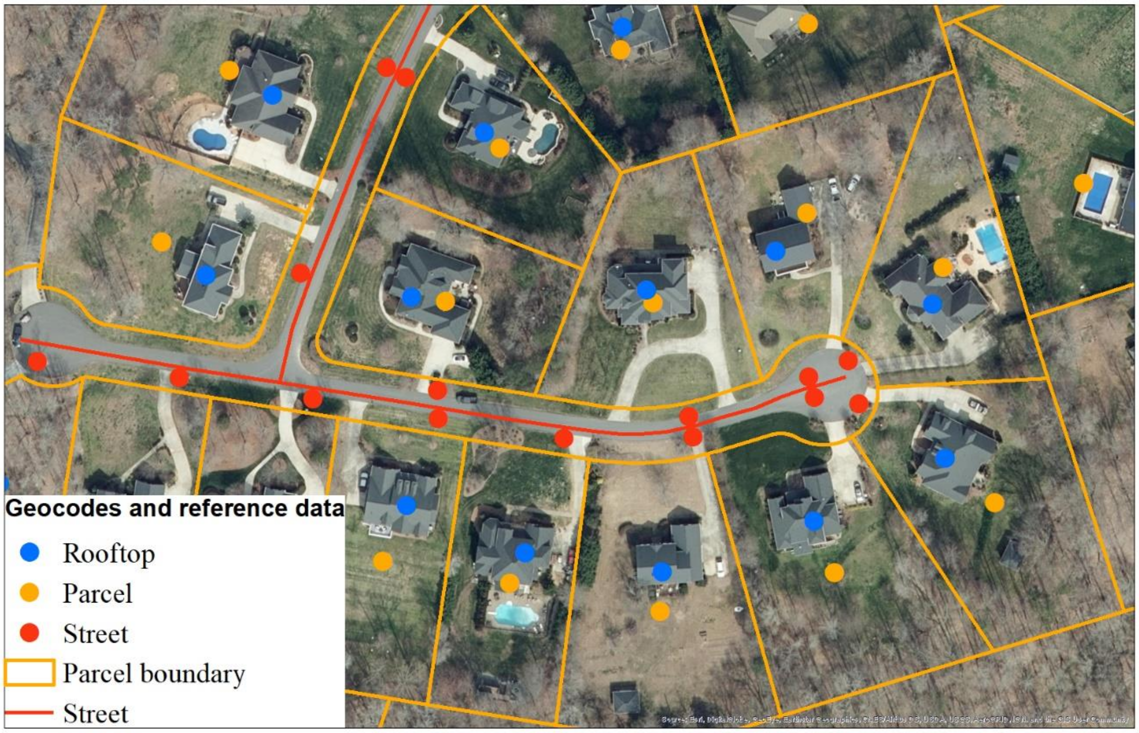

Figure 1.4: The 3 types of reference datasets: rooftop centroid in blue (most accurate), parcel centroid (reasonably accurate) and street centerline (less accurate)

Positional accuracy has been examined for cancer records (Rushton et al., 2006), locations of air traffic-related air pollution (Whitsel et al., 2004, Zandbergen, 2007), and sex-offenders records (Zandbergen and Hart 2009). Mazumdar et al. (2008) found that disease rates were poorly characterized when street geocodes were used compared to rooftop geocodes. Jacquez (2012) further iterates that differences in positional accuracy for different geocoding techniques in environmental health analysis lead to exposure mischaracterization. This can affect the reliability of spatial analysis and modeling estimates which are critical in decision making (Cressie and Kornak 2003; Zandbergen 2009). Presently, there are few literatures on predicting geocoding errors for private wells. This gap in knowledge is critical since the magnitude and variance of geocoding positional accuracy has also been found to vary geographically, with urban areas having better accuracy than in rural areas (Cayo and Talbot 2003; Zimmerman and Li 2010; Rosu and Chen 2016).

Although, Owusu et al. (2017) examined positional accuracy of private wells in Gaston County, and found that positional accuracy was high for parcel geocodes compared to street geocodes, the analysis was for only 287 wells and the true location was assumed to be rooftop geocodes. By the time of this dissertation GPS coordinates were used as the true location to determine positional accuracy of rooftop, parcel and street geocodes for 1075 private wells (sample size in October 2019) in Gaston County. The findings from this study can inform the choice of reference data to use to improve positional accuracy and contributes new evidence of uncertainty in GIScience.

1.9 Geographic Approaches in Environmental Hazard Assessments

In environmental hazards assessment, geographic methods can enhance understanding on areas at risk, exposure populations and hazard sources. Geographic approaches for estimating levels of environmental hazards often involve spatial modeling and geostatistics (Gaus et al. 2003; Goovaerts et al. 2005; Meliker et al. 2008; Kim et al. 2011; Dummer et al. 2015). For example, Sanders et al. (2012) used a spatial model to estimate arsenic concentrations at the county-level in North Carolina. When the arsenic concentration is reported below a detection limit or reporting limit (e.g., < 5 µg/L) many studies have utilized geostatistical approach such as indicator kriging to estimate the occurrence of arsenic (Goovaerts et al. 2005; Lee et al. 2008; Goovaerts 2009; Hassan and Atkins 2011; Antunes and Albuquerque 2013). However, this approach does not incorporate confounding factors.

Some studies have utilized logistic regression with confounding factors (geology, well characteristics, hydrochemical factors) to predict the presence of arsenic at or above the reporting limit (Ayotte et al. 2006; VanDerwerker et al. 2018), but ordinary logistic regression does not account for spatial effects. Because arsenic values in an area may vary, ignoring spatial effects may result in a biased and under-performing model (Bo et al. 2014). Autologistic regression could be used to alleviate this problem (Griffith 2004; Dormann 2007; Fu et al. 2013; Bo et al. 2014).

An autologistic regression is a model that incorporates spatial autocorrelation (autocovariate) variable into a logistic regression model to obtain robust inference of the dependent variable (Griffith 2004; Dormann 2007; Fu et al. 2013; Bo et al. 2014). The autocovariate variable introduced in an autologistic regression reflects the first law of geography, suggesting near things are more related than distant things (Tobler 1970; Tobler 1979; Miller 2004). Although the autologistic regression has been applied to study distribution in plant species (Wu and Huffer 1997), hand, foot and mouth disease (Bo et al. 2014), we did not find a study of any environmental contaminant. In this dissertation, the autologistic regression was used to predict the presence of arsenic at or above detectable limits (≥ 5 µg/L).

Furthermore, in this dissertation, a non-spatial multivariate logistic regression model would be used to predict the probability of having coliform bacteria in wells. GIS maps would also be used to supplement the results of the multivariate logistic regression modeling to understand the extent of the problem of water contamination in private wells which might have been disjointed if only one approach was used. Overall, the contributions of this dissertation to the field of GIScience and health geography are presented in three research objectives discussed in the next section.

1.10 Research Objectives

The dissertation emanates from a broader project (Healthy Wells) to develop an accessible digital database of private wells for Gaston County, North Carolina. Free water sampling for coliform bacteria was administered as part of the Healthy Wells project. Results of arsenic samples were obtained from the North Carolina Department of Health and Human Services (NCDHHS), Division of State Laboratory of Public Health. The goal of this dissertation is to incorporate multilevel geocoding and spatial modeling techniques to predict the risk of arsenic and coliform bacteria in private wells. To achieve the goal, the research objectives and hypotheses are stated below.

1. Develop an enhanced approach to geocode private wells data and evaluate the positional accuracy of the geocoded data from field-collected global positioning system (GPS) coordinates.

Hypothesis 1. Using multiple reference datasets, probabilistic record linkage (data matching technique) in a multi-stage approach may not improve geocoding match rate.

Hypothesis 2. There is no variation in the positional accuracy of rooftop, parcel and street geocodes of private wells.

2. Evaluate if the geology, pH, and well depth can improve our ability to predict the presence of arsenic at or above detectable levels (≥ 5 µg/L) found in private wells using an autologistic regression model.

Hypothesis 1. The geology, pH and well depth cannot predict the likelihood of arsenic being present in wells at or above detectable levels.

Hypothesis 2. There are no discernable spatial patterns in the probability of arsenic being present in wells.

3. Identify whether the type of well, well age, well depth, parcel size, and soil ratings for a leachfield can predict coliform bacteria presence in wells using multivariate logistic regression.

Hypothesis. The type of well, well age, well depth, well, parcel size and soil rating for a leachfield cannot predict the probability of coliform bacteria being present in wells.

1.11 Contributions

This dissertation attempts to develop a novel approach to geocode addresses even when there missing or incomplete information. The findings can be adopted to increase geocoding match rate while ensuring most geocodes have good positional accuracy in the development of GIS data of private wells. The geocoding approach and GIS data were critical for creating predictive models for arsenic and coliform bacteria. The three objectives inform and complement each other – filling in knowledge gaps that would be apparent if each study was treated separately. In objective 1, a systematic approach to develop a GIS database of private wells with good positional accuracy is presented.

In objective 2, the geocoding approach developed in objective 1 is used to obtain geographic coordinates of arsenic samples for samples with no GPS coordinates. Data on geology, well depth and pH were evaluated to predict arsenic at or above detectable levels using a spatial autologistic model. This model has not been applied to study environmental hazards with many non-detects and this dissertation bridges this gap in the literature. In objective 3, using the GIS data obtained from objective 1, student teams were sent to collect water samples and test for the presence of coliform bacteria. . The percentage of positive samples with a 1 kilometer is assessed to determine spatial patterns of contamination. The use of GIS maps and multivariate logistic regression modeling in the analysis of the coliform data can complement each other to understand the extent and causes of coliform bacteria in wells.

This dissertation provides a framework to develop an accurate GIS database, while testing and evaluating the utility of spatial autologistic regression for the analysis of environmental contaminants with samples below a reporting limit. In addition, complementing multivariate logistic regression modeling with GIS maps provides an effective and informative approach to answer questions on where, why and what factors contribute to the presence of coliform bacteria in wells. Although Objectives 2 and 3 are not directly linked, they both offer a good overview of the water quality in private wells and reiterate the need for clear and accurate methods for digitizing paper records.

1.12 Outline of the Dissertation

The remainder of the dissertation is organized as a collection of three papers. Each article addresses an objective using different methods. As shown in *Figure 1.5*, the structure of the dissertation showing the major processes to achieve each objective.

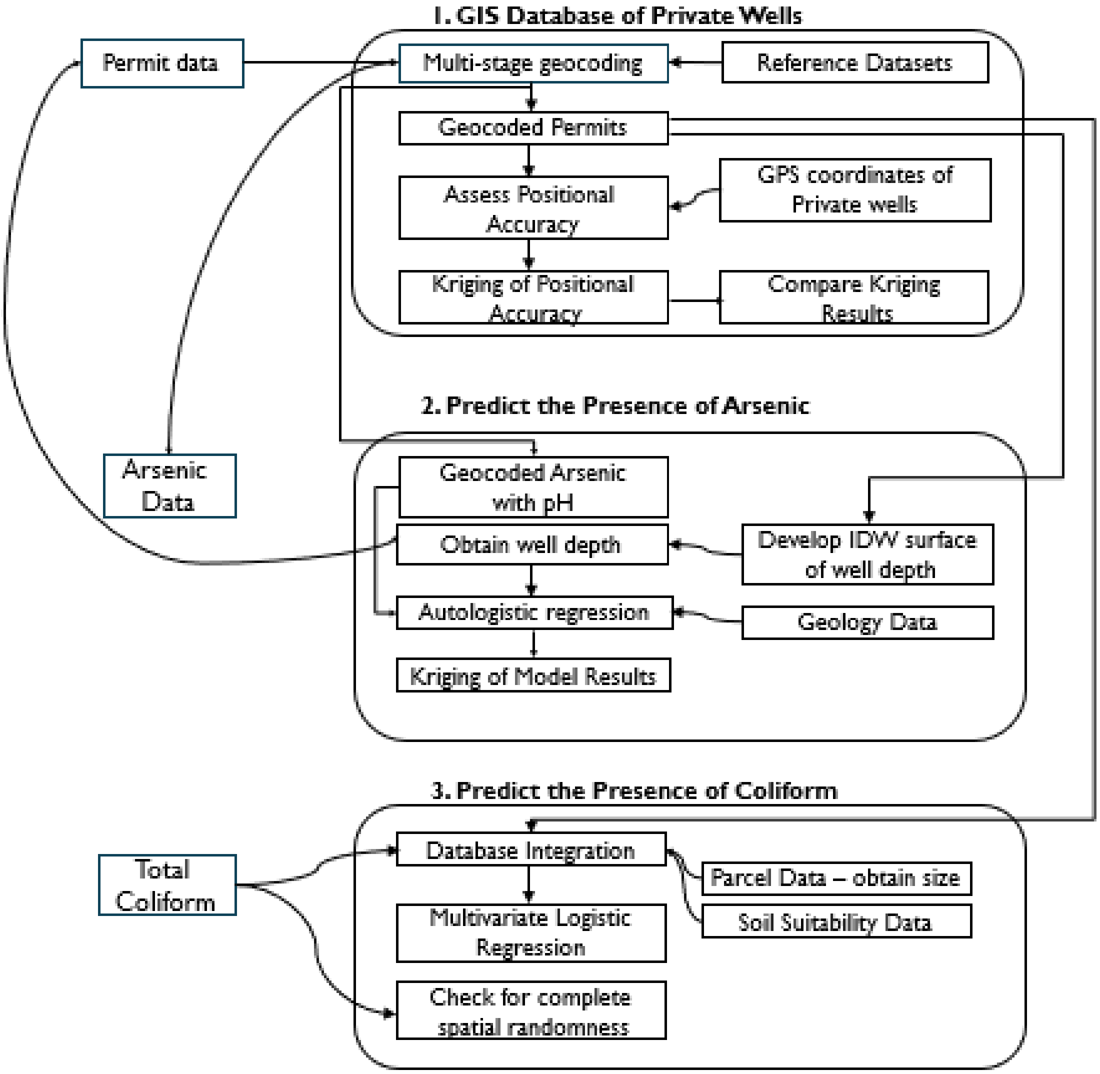
1. GIS Database of Private Wells
Permit data
Multi-stage geocoding
Reference Datasets
Geocoded Permits
Assess Positional Accuracy
GPS coordinates of Private wells
Kriging of Positional Accuracy
Compare Kriging Results
2. Predict the Presence of Arsenic
Arsenic Data
Geocoded Arsenic with pH
Obtain well depth
Develop IDW surface of well depth
Autologistic regression
Geology Data
Kriging of Model Results
3. Predict the Presence of Coliform
Total Coliform
Database Integration
Parcel Data – obtain size
Soil Suitability Data
Multivariate Logistic Regression
Check for complete spatial randomness

Figure 1.5: Workflow for the various components in the three objectives

In Chapter 2, multiple reference datasets are used to develop a multi-stage geocoding approach to geocoded permits data. Positional accuracy of the geocodes are evaluated from field collected GPS coordinates. Kriging interpolation are then used and results for rooftop, parcel and street positional accuracy are compared. The manuscript detailing the methods and results is currently under review in the Journal of Environmental Health.

In Chapter 3, the multi-stage geocoding approach developed in chapter 1 was used to extract geographic coordinates for arsenic data. Well depth information is then merged from the permit data, but for those with missing depth, the depths are estimated from an interpolation surface. Next, the geology information is appended to the arsenic samples which also have pH and well depth information. Then, the geology, well depth, and pH are examined using autologistic regression model to predict the probability of arsenic being present at or above detectable levels. Model results are examined for spatial patterns using kriging. The manuscript detailing the methods and results is under review in the Journal of Health and Exposure.

In Chapter 4, coliform bacteria samples are merged with well attributes (well age, well depth, type of well), soil ratings for a leachfield, and parcels. This data serves as input parameters to predict the probability of coliform being present using multivariate logistic regression Chapter 5 provides the overall conclusion from the three manuscripts, how the studies are related, recommendations, and future research directions.

CHAPTER 2: A MULTI-STAGE GEOCODING APPROACH FOR THE DEVELOPMENT OF PRIVATE WELLS DATABASE, GASTON COUNTY, NORTH CAROLINA[1]

Abstract

Many existing inventories of private wells in the United States lack digital geographic coordinates, and county-level permitting system often store information in paper copies. We developed a GIS database of private wells from paper permits issued since 1989 in Gaston County, North Carolina (n = 8,721) using a multi-stage geocoding approach. We then assessed the positional accuracy of the geocodes from the field-collected GPS location of these wells. In total, 98.9% of permits were successfully geocoded, and 12.3% were secured with GPS devices. There were significant differences ($p < 0.05$) in positional accuracy for rooftop, parcel, and street geocodes of private wells in the GIS database, but positional accuracy was high for rooftop geocodes. Our approach is portable to other regions interested in the development of a digital private wells inventory digitally with GIS to aid in monitoring water quality and planning accurate public health interventions.

Keywords: Private well, geocoding, GIS, database, positional accuracy

[1] Owusu, C., Delmelle, E., Tang, W., Silverman G.S, Dye, S. Under review in *Journal of Environmental Health*

2.1 Introduction

Although most Americans drink water from public water systems, about 13 million households (45 million Americans) rely on unregulated private wells for drinking water (United States Environmental Protection Agency, 2019). Private wells use groundwater which is prone to contamination. Sources of pollution to groundwater include leaks from coal ash ponds (Huggins, Senior, Chu, Ladwig, & Huffman, 2007), underground storage tanks (Fabro, Ávila, Alberich, Sansores, & Camargo-Valero, 2015), landfills, septic systems (Schaider, Ackerman, & Rudel, 2016), and excessive fertilizer application and animal waste (Messier, Kane, Bolich, & Serre, 2014). Groundwater contamination can also originate from native rocks. For example, the native rocks in the Piedmont region, including Gaston County, North Carolina, is associated with high levels of arsenic in groundwater (Harden, Chapman, & Harned, 2009; Pippin, 2005).

In North Carolina, local health departments issue permits, and test for bacteria and inorganic chemical contaminants after the construction of a new private well (MacDonald Gibson & Pieper, 2017). Copies of the permits are kept on file by most counties in paper formats. For example, Mecklenburg County (Mecklenburg County Health Department, 2019) has developed digital geographic information system (GIS) database showing the geographic locations of their private wells, but for many other counties this is unavailable. The lack of geographic coordinates of private wells pose challenges in modeling exposure to contaminants with GIS and communicating risk to well users.

In order to develop a GIS database of private wells, geocoding techniques can be used to convert location information in the form of addresses into geographic coordinates of longitude and latitude (Owusu, Lan, Zheng, Tang, & Delmelle, 2017). Geocoding an address requires spatially explicit reference datasets of road, parcels, or rooftop (i.e., rooftop centroid) to convert the address information to longitude and latitude of the reference data. These reference datasets also define the three-geocoding techniques available in GIS.

Two measures of geocoding data quality that are well recognized in the literature are match rate and positional accuracy (Goldberg, Wilson, & Knoblock, 2007; Zhan, Brender, Lima, Suarez, & Langlois, 2006). Geocoding match rate is the number of successful matched results, and it depends on the availability of up-to-date reference data (Goldberg et al., 2013). An approach to improve geocoding match rates is to combine multiple reference datasets and use hierarchical rules in a multi-stage approach. For instance, multi-stage geocoding using street and parcel datasets improved geocoding match rate of sex offenders in Hamilton, Ohio, to 90 percent (Murray, Grubesic, Wei, & Mack, 2011). Sonderman et al. (2012) incorporated multiple base-references from commercial vendors and United States Postal Services address points reference data to improve geocoding match rate to 99 percent. The geocodes can also be placed manually when nearby features are known (Goldberg, 2011; McDonald, Schwind, Goldberg, Lampley, & Wheeler, 2017). However, this approach is time-consuming.

Most studies employing geocoding suffer from incomplete or missing input data. A data-matching algorithm called probabilistic record linkage (PRL) can be used to re-engineer addresses, as long as secondary information (e.g., name, parcel number) of the records are provided. PRL is used to match two datasets with similar attributes by assigning weights based on the degree of similarity (Randall, Ferrante, Boyd, & Semmens, 2013). High weights suggest a higher probability of a match (Schmidlin, Clough-Gorr, & Spoerri, 2015). PRL has been used in health services research of birth outcomes and hospitalization records (Bentley, Ford, Taylor, Irvine, & Roberts, 2012), but the utility to improve geocoding match rates by re-engineering residential addresses has not yet been tested or evaluated. The application of PRL to improve match rates is critical since excluding non-geocoded records will likely reduce sample size and weakens the generalization of the analytical results due to selection bias (Ha et al., 2016; Zandbergen, 2009; Zimmerman, 2008).

Another metric for the quality of geocoding results is positional accuracy, and it refers to the distance between the position of the geocode and its true location (Bonner et al., 2003; Ward et al., 2005). The smaller the error distance, the higher the accuracy of the geocode. Differences in positional accuracy in environmental health assessments may lead to exposure mischaracterization and can affect the reliability of spatial modeling estimates (Zandbergen, 2009). For instance, when geocoded data of contaminated private wells are analyzed, larger error distances can affect the estimate for contaminants characterized by small ranges beyond which spatial autocorrelation vanishes.

In this article, we describe a multi-stage geocoding approach used to develop a GIS database of private wells. We then assess the positional accuracy of the geocodes in our GIS database using field-collected GPS locations of private wells. The study provides a novel approach to increase geocoding match rate that goes beyond using multiple reference datasets to implementing PRL technique. The approach is portable to other counties in need of a digital database of private wells to aid in spatial modeling of exposure to contaminants, monitoring water quality, and planning public health interventions.

2.2 Study Area and Data

Private wells data were retrieved from Gaston County, North Carolina (Figure 2.1). Since 1989, a total of 8,721 permits have been issued by the Gaston County Department of Health & Human Services (GC-DHHS). A typical permit contains a unique permit number, information on the well owner, type of well, size, depth, casing depth, residential address, parcel tax location codes, and site sketch of the well. Some of the historical paper permits were particularly challenging to digitize and geocode due to damaged paper, missing address information, directional addresses, and illegible handwriting.

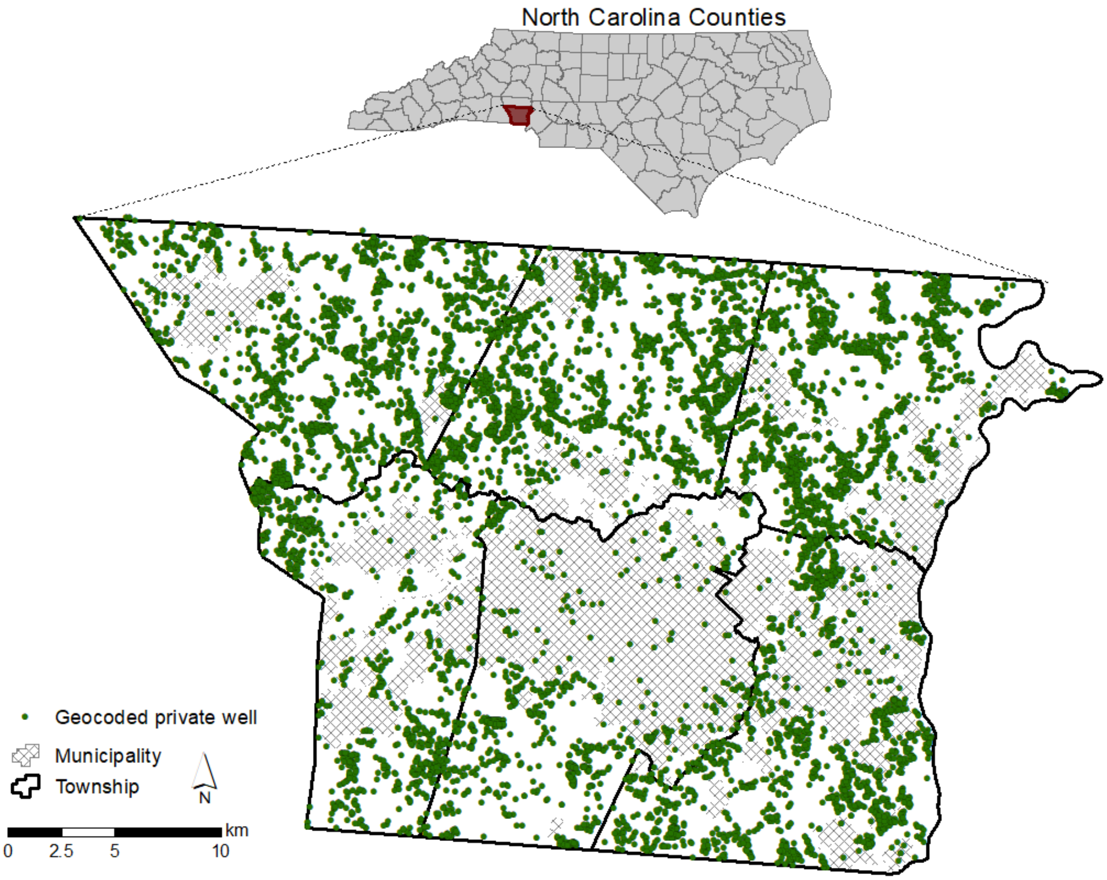

Figure 2.1: a) Location of geocoded private wells in Gaston County (Geocoding approach described in the methodology section)

2.3 Reference Data

Spatially explicit reference datasets for rooftop centroid, parcels, and road networks from Gaston County were retrieved from a variety of sources (Table 2.1). Other non-spatial reference data included deed records, property tax information, paper copies of laboratory test results for total coliform, and inorganic chemical laboratory test results. The paper copies of total coliform and inorganic chemical laboratory test results were limited to private wells since 2008 when North Carolina mandated laboratory testing once a new well is constructed (North Carolina Department of Health and Human Services, 2019).

Table 2.1: Reference data and source used in this study

Reference Data	Year	Source
Rooftop centroid data	2016	Gaston County IT-GIS Department
Parcel data	2012, 2014, 2015, 2016	Gaston County Department of Planning & Development Service
County roads	2002, 2016	Gaston County IT & GIS Department
Tigerline roads	1992, 2000	U.S. Census Bureau
Copies of Coliform and Inorganic chemical test results	2008 - 2016	Gaston County Dept. of Health & Human Service
Deeds records	2012, 2014, 2015, 2016	Gaston County Dept. of Planning & Development Service
Property tax information	2012, 2014, 2015, 2016	Gaston County Dept. of Planning & Development Service

2.4 Parsing and Address Cleaning

Incomplete address information and lack of address standardization may prohibit geocoding automation (Rushton et al. 2006; Goldberg 2011; Murray et al. 2011; Sonderman et al. 2012; Rosu and Chen 2016). Initial steps were taken to standardize and evaluate the raw addresses. The raw addresses were parsed into usable components, including street number, prefix direction (e.g., “S”), street name, type, and suffix direction (e.g., “SE”) when available. Common data entry errors such as (e.g., STRET”, “CIRCL”) were corrected. Other manual data cleaning strategies such as sorting and filtering by common street names helped correct typographical errors. However, more than half of the input permit addresses were postal box entries (e.g., “PO Box 101), missing, or directional descriptions.

2.5 Methodology

2.5.1 Multi-Stage Geocoding

Due to the inherent uncertainty of some input permit addresses, we developed a multi-stage geocoding to increase the number of successfully geocoded private wells. The major components were the input permit and reference data, the geocoding procedure, and output geocoded private wells (Figure 2.1). The geocoding procedure consisted of two stages: automation and improvement.

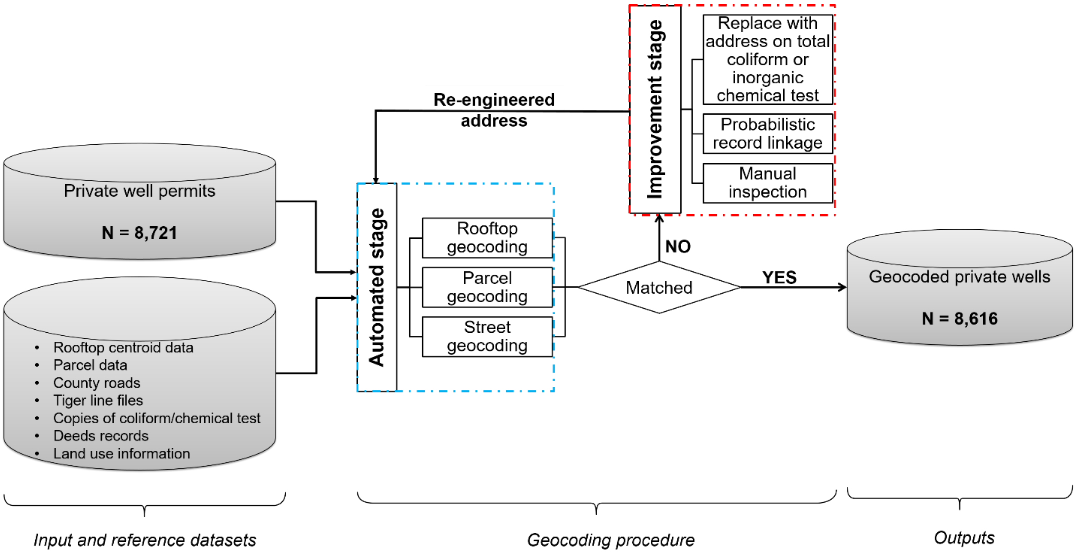

Figure 2.2: The multi-stage geocoding workflow for private well permits in Gaston County, NC

2.5.2 Automation Stage

During the automation stage, rooftop, parcel, and road geocoding techniques were combined hierarchically based on their spatiotemporal accuracy into a composite address locator (using ESRI ArcGIS 10.6). An address locator is a model used to create geometry for input addresses during geocoding. In the composite address locator, the input permit addresses would first attempt to geocode them with rooftop geocoding. Unsuccessful input addresses were then considered at the parcel geocoding level. However, when the input addresses were not geocoded in parcel geocoding, they were considered for street geocoding. The input permit addresses that were not geocoded after the first geocoding trial were considered in the improvement stage (Figure 2.2).

2.5.3 Improvement Stage

The first approach in the improvement stage was to replace the missing/incomplete permit addresses with re-engineered addresses from copies of total coliform or inorganic chemical tests that were linked using the unique permit numbers. The new re-engineered addresses were then transferred to the automation stage to be geocoded using the composite address locator. Permits not geocoded were then considered for the PRL approach.

PRL data matching technique was implemented (using LinkageWiz ^{TM} 2016) as a second approach in the improvement stage to link the permit data (source) with the parcel data (reference). Parcel attributes such as tax location codes, parcel owner information (surname, middle name, first name), parcel street name, parcel size, lot number, and subdivision name were paired with corresponding attributes available on the private well permits. PRL results are evaluated by the weight scores associated with a potential match based on the field agreements, disagreements, and missing values during the linkage (Bentley, Ford, Taylor, Irvine, & Roberts, 2012; Randall, Ferrante, Boyd, & Semmens, 2013; Schmidlin, Clough-Gorr, & Spoerri, 2015). The weights are derived from the logarithm of the frequency ratio of the common attributes being examined and is expressed as;

$$Weight = \log_2\left(\frac{\text{Frequency of agreement in LINKED pairs } in\; s_1, r_1}{\text{Frequency of agreement in UNLINKED pairs in } in\; s_1, r_1}\right) \qquad (1)$$

with s_1, r_1 as the common attributes in the permit and parcel data, respectively.

The total weight scores were evaluated for potential and false-positive linkages. A higher value corresponds to a good potential match, while a low value may signal a false match. We accepted linkage pairs with weight scores above 30 because there was a natural break in the distribution of weights beyond the score, but manually reviewed those below this score before a potential linkage was accepted. The corresponding addresses associated with the parcels were used to replace the incomplete or directional addresses in the permit data. Re-engineered addresses were transferred back to the automation stage. Permits not geocoded were transferred to the final approach in the improvement stage.

The final approach was to manually inspect only the permits that were not geocoded after PRL by comparing them with information in deeds, and parcel data to trace any record of change in ownership that could help identify the addresses on these permits. Once a permit was found to have corresponding information in the deed or parcel data during the manual inspection, the incomplete or directional address on the permit was replaced accordingly with the address in the deed or parcel data. The new re-engineered addresses were then transferred to the automation stage to be geocoded using the composite address locator. Permit addresses not geocoded after this approach were excluded from the GIS database of private wells.

2.5.4 Field Data Collection of GPS Coordinates of Private Well

We organized students from University of North Carolina at Charlotte (UNC Charlotte) into a two-member team to get the coordinates at the actual well sites (using Mesa ^{TM} handheld GPS) and compare them to the geocoded locations. In order to minimize the drive-time to the locations, we used the database and GIS network analysis to develop optimized route schedules for each team. In some instances, student teams also sampled private wells along a street when the route schedules were not appropriate. The county's environmental health department provided the necessary training to the students, and in collaboration with UNC Charlotte coordinated the field data collection of the private well locations, and also offered free water sampling for total coliform. Using the unique permit identification numbers assigned when it is issued, the field determined GPS coordinates data were merged to the GIS database of private wells.

2.5.5 Assessing Positional Accuracy

The error distance between the field measured GPS coordinates and geocodes obtained from rooftop, parcel, and street geocoding of the private wells were calculated as an indicator of positional accuracy. Because only 12.3% of well owners agreed to securing the GPS coordinates of their private well, for this study, we used kriging interpolation techniques to estimate the positional accuracy of the geocodes at unsampled locations. Kriging is based on the geostatistical theory of regionalized variables, which states that variables in an area exhibit both random and spatially structured properties (Goovaerts, 2000; Pyrcz & Deutsch, 2014). The resulting interpolated surface obtained from kriging error distances were mapped and visually compared to ascertain the geographic variation in positional accuracy of rooftop, parcel, and street geocodes.

The skewed error distances for rooftop, parcel for street geocoded results were log-transformed before kriging so the data were normally distributed, and the results later back-transformed for interpretation purposes. We fitted the kriging semivariograms with an exponential model because it yielded the smaller sum of squared errors for rooftop and parcel error distances. Although the sum of squared errors for street error distances was small for Gaussian model fitting, we used an exponential model in order to compare the predictions for rooftop, parcel, and street geocoding.

2.6 Results

From October 2016 to September 2019, a total of 8,721 permits were digitized. Only 3,207 (38.0%) of these permits were geocoded automatically at the first attempt. The remaining permits had incomplete or missing address information. The improvement stage approaches to re-engineer permit addresses yielded an additional 5,298 (60.7%) geocodes in the private wells GIS database. Individually, PRL added more re-engineered addresses (2,054 (23.6%)) compared to substituting addresses on copies of laboratory reports of coliform or inorganic chemical tests (1,917 (22.0%)) and manual inspection interventions (1,327 (15.2%)) (Table 2.2). A total of 105 (1.1%) of the private well permits were not geocoded because of paper damage, illegible handwriting, or directional addresses. Geocoded permits were more likely to be at the rooftop level (92.3%) compared to parcel (3.5%) and street level (3.1%).

Table 2.2: Summary of results for different geocoding stages.

Stage	Rooftop geocoding	Parcel geocoding	Street geocoding	Total
	n (%)	n (%)	n (%)	n (%)
Original permit	3115 (35.7)	46 (0.5)	157 (1.8)	3318 (38.0)
Re-engineered addresses				
From copies of coliform/inorganic chemical test	1824 (20.9)	14 (0.2)	79 (0.9)	1917 (22.0)
PRL	1875 (21.5)	179 (2.1)	--	2054 (23.6)

Manual inspection	1235 (14.2)	59 (0.7)	33 (0.4)	1327 (15.3)

Total match = 8,616 (98.9)

PRL = Probabilistic record linkage

2.6.1 Positional Accuracy

From October 2017 to September 2019, 1,075 households agreed to have their private well locations secured with GPS. The field teams reported reasons for not securing the GPS coordinates as: 1) owners were not at home (n = 3,877); 2) property could not be entered because of notices such as no trespassing, beware of dogs or private property (n = 2,039); 3) residents declined to participate (n =1,083); 4) residents were at home but unavailable, residents asked to have their well locations taken later (n = 240); house was serviced by city water (n = 107) and house was on community well (n = 31). A total of 164 well owners requested not to have their well locations identified by GPS.

Positional accuracy was statistically different ($p < 0.05$) for rooftop, parcel, and street geocodes. The rooftop was best (mean = 26m, standard deviation = 15 m), followed by parcel (mean = 44m, standard deviation = 44 m), and then street geocodes (mean = 72m, standard deviation = 61 m). Median positional accuracy for rooftop, parcel, and street geocoding were 24m, 32m and 52m, respectively. The cumulative frequency distribution (Figure 2.3) shows that 95% of the rooftop positional accuracy was within 52 meters, with 95% of parcel positional accuracy within 130m, and 95% of the street positional accuracy within 190m.

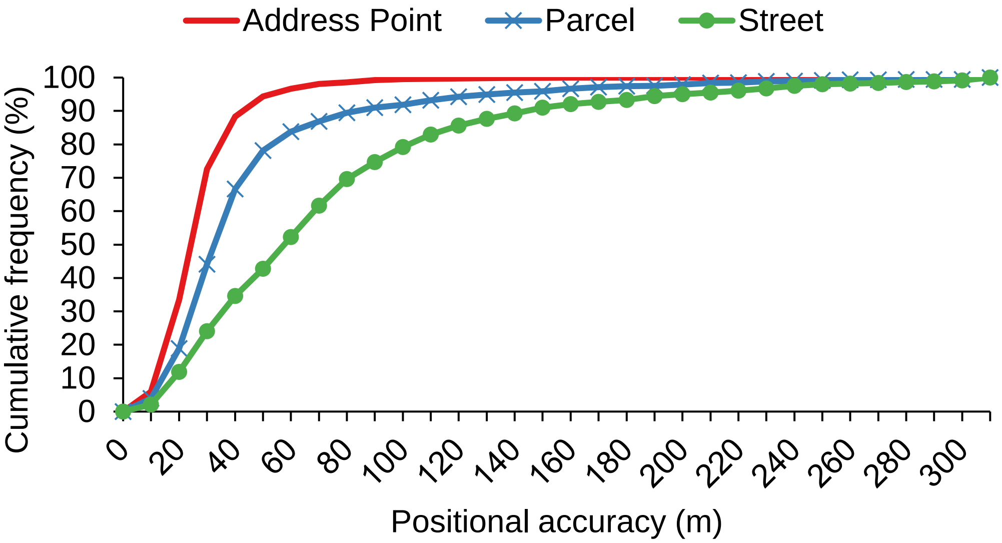

Figure 2.3: Cumulative distribution of positional accuracy (in meters) of rooftop, parcel, and street geocoded locations of private wells to GPS measured positions (n = 1,075).

The kriging maps for rooftop geocodes (Figure 2.4a), parcel geocodes (Figure 2.4b), and street geocodes (Figure 2.4c) show geographic variations indicating differences in positional accuracy across the county. For example, whereas the map for rooftop shows that positional accuracy does not exceed 60m, the parcel kriging map shows that some parts of the county have positional accuracy greater than or equal to 120m. Cross-examination of parcel sizes in these sections revealed that the area has larger parcel sizes (mean = 28,059 m^2) than in other parts of the county. Nonetheless, the kriging map for street positional accuracy shows more areas with values greater than or equal to 120m.

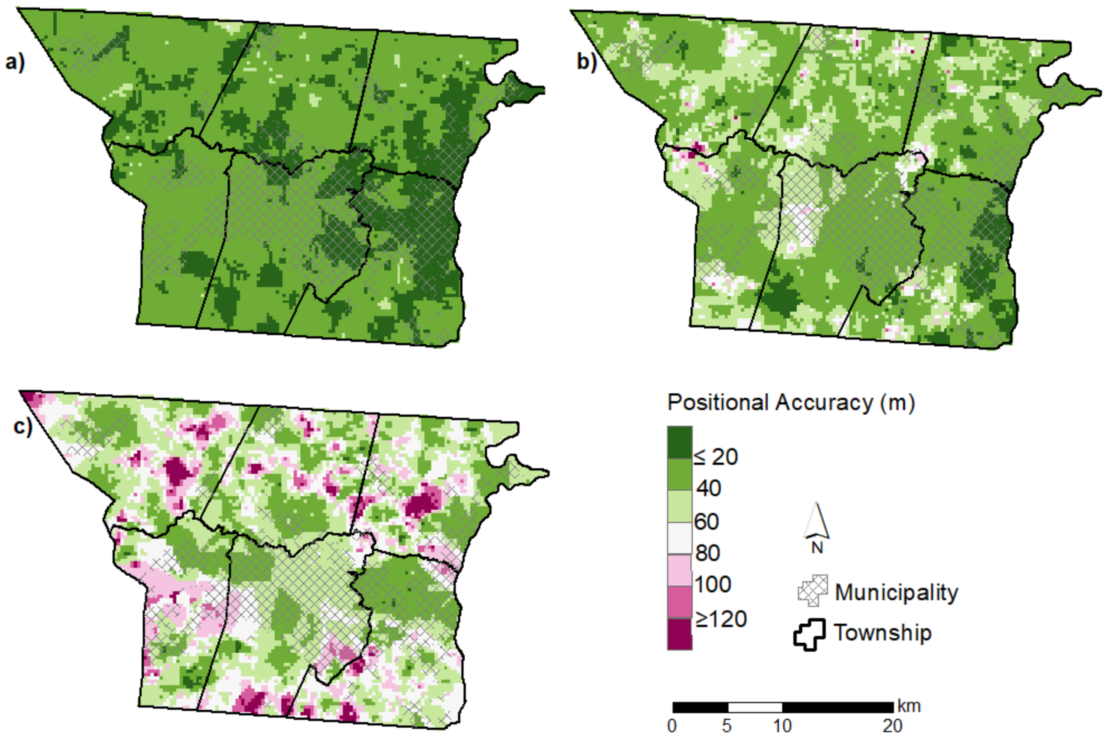

Figure 2.4: Kriging-based prediction of private well positional accuracy from geocodes obtained from: a) rooftop, b) parcel, and c) street geocoding techniques in the GIS database, Gaston County.

2.7 Discussion

Developing a GIS database of private wells presented several challenges; the key among them was dealing with paper permits and improving the geocoding match rate from permits with missing or incomplete address. This was resolved by identifying missing addresses from other data sources such as laboratory test results. Additionally, our PRL approach connected information such as tax location codes on private wells permits with GIS parcel data to help identify missing addresses. These approaches improved our geocoding match rate from 38.0% to 98.9%. Our high match rate helps show the potential of this approach to other institutions interested in developing a GIS database of private wells.

Although we could not geocode 105 (1.1%) permits into the GIS database, the high geocoding match rate was enough to meet the health department's needs of showing the geographic locations of their private wells. Information on coliform or chemical measurements at the well can be integrated into the GIS database to aid in monitoring water quality. For example, the developed digital database of private wells allows identification of potentially contaminated wells from an easily accessible dataset.

Our positional accuracy assessment shows that rooftop geocoding outperforms parcel and street geocoding in the spatial representation of a private well. This may be because rooftop geocoding outputs are to the centroid of a building rooftop, and private wells are mostly constructed near the residence. In environmental exposure assessment, GPS coordinates serve as the best spatial data in modeling exposure to the private well (Cromley & McLafferty, 2011). However, when the GPS coordinates are unavailable, rooftop geocodes can be used.

A limitation of the multi-stage geocoding approach employed in this study may result from inputting inaccurate addresses from existing permits. For example, data entry errors from the paper permits could have a significant impact on the final geocoded data. The PRL approach also is limited by the availability of common attributes in the permit and parcel data. Even though PRL application is popular in health services research, privacy issues may arise from linking data with sensitive information (Schmidlin, Clough-Gorr, & Spoerri, 2015). In this study, only publicly available parcel data were used as a reference during the linkage. The major strength of our approach was the use of multiple reference datasets and techniques to augment the geocoding rate of permit addresses in the development of a GIS database of private wells.

2.8 Conclusions

An accurate private well GIS database is critical to evaluations of local environmental factors associated with groundwater contamination and threats to rural drinking water quality. We were successful in creating such a database for Gaston County, North Carolina, using a multi-stage technique that should be practical for many health departments or other agencies currently limited to paper file records of domestic wells. We built our database first from paper files that often were damaged, missing address information, using directional addresses, or illegible. We supplemented this initial information by obtaining addresses from well water testing done on a limited basis by the State laboratory. Lastly, for those locations for which reliable addresses were still unavailable, we employed probabilistic record linkage, and used three distinct geocoding databases: rooftop, parcel, and street. An important finding was that rooftop geocoding was the most accurate technique, negating the need for using the other two databases for most foreseeable applications.

Potential applications of the database include the production of maps showing the locations of private wells and identifying the relative risk of contamination on an areal basis. This information could be useful to support decision-making, for instance, when to alert county residents of the risks posed by contaminated groundwater in their vicinity, or where to conduct additional sampling in areas deemed at risk. A variety of other public health interventions could be facilitated through the spatial identification of local wells, and the accessibility of digitally available data.

CHAPTER 3: A SPATIAL AUTOLOGISTIC MODEL TO PREDICT THE PRESENCE OF ARSENIC IN PRIVATE WELLS ACROSS GASTON COUNTY, NORTH CAROLINA USING GEOLOGY, WELL-DEPTH, AND PH[2]

Abstract

Chronic exposure to arsenic-contaminated drinking water is detrimental to human health. We develop an autologistic regression model to evaluate if the geology, pH, and well depth can improve our ability to predict the presence of arsenic at and above detectable levels (≥ 5 µg/L) found in private wells. We use arsenic samples measured in private well water across Gaston County, North Carolina, from 2011 to 2017. We use kriging to map the probability of arsenic at detectable levels across Gaston County. Arsenic at detectable levels was reported at 78 private wells. The median pH for samples containing detectable levels of arsenic was 7.3 and for samples with arsenic < 5 µg/L was 7.1. Our spatial autologistic model suggests that arsenic at detectable levels is positively associated with pH. In addition, private wells set in Mica schist (€Zms) were associated with arsenic, suggesting a local-scale geologic source influence of arsenic in the county. Our kriging map shows that the northwestern section of the county has more than a 50 percent probability to have arsenic at detectable levels. In conclusion, based on our results, we recommend increased testing for wells in the Mica schist area. The map of probability of arsenic at and above detectable levels can be used to implement cost-effective targeted interventions.

Keywords: *Arsenic, Autologistic Regression, Geology, GIS**,** Private Wells, Water*

[2] Owusu, C., Silverman, G.S, Vinson, D., Bobyarchick, A., Paul, R., Delmelle, E. (2020) *Journal of Health and Exposure. https://doi.org/10.1007/s12403-020-00373-6*

3.1 Background and Rationale

Chronic exposure to elevated arsenic levels (>10 µg/L) in drinking water has been associated with several types of cancers including prostate (Benbrahim-Tallaa and Waalkes 2008), lung (Heck et al. 2009; Dauphiné et al. 2013), bladder (Steinmaus et al. 2003), kidney (Yuan et al. 2010), and skin (Karagas et al. 2015). Recent studies have suggested that even low levels of arsenic (<10 µg/L) in drinking water may impact fetal development (Bloom et al. 2016; Almberg et al. 2017), increase odds of diabetes (Mahram et al. 2013), and cause an elevated risk of heart disease (Bräuner et al. 2014; James et al. 2015).

In the United States (U.S.) alone, among the 44.1 million Americans using private wells, 2.1 million individuals use groundwater above the drinking water arsenic standard of 10 µg/L set by the U.S. Environmental Protection Agency (USEPA) (Ayotte et al. 2017). Yet, private wells are not regulated in the U.S (MacDonald Gibson and Pieper 2017). In Gaston County, North Carolina (the focus of our study; Figure 3.1), nearly 42% of residents rely on private well water (Centers for Disease Control and Prevention (CDC) 2019). The accurate prediction of the spatial and/or vertical variation of arsenic in groundwater systems is critical to water supply management.

Arsenic has been found at elevated levels in aquifers across North Carolina, US (Sanders et al. 2012), China (He et al. 2020a), Bangladesh (Hossain and Sivakumar 2006), Nepal (Gurung et al. 2005) and many other regions (Smedley and Kinniburgh 2012). In North Carolina, some studies have underlined possible associations between elevated arsenic concentrations and metavolcanic or metavolcaniclastic rocks (Pippin 2005; Harden et al. 2009), and metamorphosed clastic sedimentary rocks (Chapman et al.

2013). Reid et al. (2005) have suggested that the occurrence of arsenic in groundwater in the Piedmont of North Carolina could be related to fracture coatings in iron-manganese filled borehole cores from oxidized zones. The northwestern part of Gaston County, North Carolina (Figure 3.13.1) is within the area described as the physiographic and general geologic Piedmont of North Carolina. Chapman et al. (2013) suggested that elevated arsenic concentrations in groundwater from rock units are positively correlated with pH of 7.2 or greater in the Piedmont of North Carolina. At elevated pH, the formation of soluble oxyanions can increase arsenic mobilization through desorption processes (Ayotte et al. 2003; 2006).

Most private wells in the Piedmont of North Carolina obtain water by drilling into bedrock, but a few wells tap water from the regolith at shallow depth (Daniel and Dahlen 2002). Two studies have examined the relationship between arsenic concentration and well depth in bedrock aquifers in the Piedmont of North Carolina. Kim et al. (2011) found associations between elevated arsenic levels and deep wells within welded tuffs and quartz units that were close to the transition zones between primarily pyroclastic and primarily volcaniclastic sedimentary rocks. Chapman et al. (2013) found that arsenic concentrations in crystalline lithologies were positively correlated with well depth. However, to date there is no predictive model of the presence of arsenic to guide well planning in Gaston County, even though the county-level analysis suggests most private wells exceed the U.S. EPA drinking water standards for arsenic (Sanders et al. 2012). The complex spatial distribution of geologic formations make it difficult to assume that arsenic concentrations would be evenly distributed in the county.

Spatial modeling and geostatistics have received considerable attention for the prediction of arsenic in groundwater (Gaus et al. 2003; Goovaerts et al. 2005; Meliker et al. 2008; Kim et al. 2011; Dummer et al. 2015). When most of the data contain arsenic values that are reported as below the reporting limit, researchers have relied on geostatistical techniques such as indicator kriging to estimate the occurrence of arsenic (Goovaerts et al. 2005; Lee et al. 2008; Goovaerts 2009; Hassan and Atkins 2011; Antunes and Albuquerque 2013), yet these approaches typically do not incorporate predictor variables. Some studies have used logistic regression with various predictors (geologic and anthropogenic sources of arsenic, geochemical processes, hydrogeologic, and land-use factors) to model the occurrence of arsenic ≥ 5 µg/L (Ayotte et al. 2006; Bretzler et al. 2017). The ordinary logistic regression is based on the assumption that the relationship between the presence of arsenic and potential confounding factors would not change across a region. However, spatial autocorrelation, defined as a measure of the similarity in values for nearby observations (Griffith 1987), is frequently present in environmental data. For example, there are different geologic regions in Gaston County, and samples taken from those distinct regions may exhibit strong similarities, violating the assumption of spatial stationarity. Therefore, ignoring spatial effects in ordinary logistic regression could result in a biased and under-performing model (Bo et al. 2014). Autologistic regression could be used to alleviate this problem (Griffith 2004; Dormann 2007; Fu et al. 2013; Bo et al. 2014; Seeley et al. 2019).

The autologistic regression is a spatial model that incorporates a spatial autocorrelation (autocovariate) variable into a logistic regression model to obtain robust inference (Griffith 2004; Dormann 2007; Fu et al. 2013; Bo et al. 2014; Liu et al. 2018).

The autocovariate variable introduced in an autologistic regression reflects the first law of geography that near things are more related than distant things (Tobler 1970; Tobler 1979; Miller 2004). In this study, we assumed that the probability of arsenic occurrence in a private well is higher if it is also present in nearby private wells, because wells located in the same rock type are more likely to have a similar elevated arsenic than comparison with a randomly selected well. The autologistic regression has gained attention in ecological studies (Wu and Huffer 1997; Dormann 2007; Tsuyuki 2008), transportation research (Liu and Sharma 2019), but have not been applied to model the occurrence of arsenic.

We develop a spatial autologistic regression model to evaluate if the geology, pH, and well depth can improve our ability to predict the presence of arsenic at and above detectable levels (≥5 µg/L) in private wells. We used this threshold because all arsenic concentration data in our study used EPA method 200.8 that has a detection limit of 5 µg/L (USEPA 1994; North Carolina Department of Health and Human Services 2020). Also, we used this threshold because lifetime exposure to even relatively low arsenic concentration can have adverse health effects.

3.2 Study Area and Geologic Setting

Gaston County, North Carolina (364 mi^2 or 942km^2) is a fast-growing county of nearly 225,000 residents (2019) in the South-Central Piedmont section of North Carolina. The county is bounded on the east by the Catawba River and Mecklenburg County, on the west by Cleveland County, on the north by Lincoln County and on the south by York County, South Carolina. Gaston County is characterized by a temperate climate with moderate temperature variations and humidity. The topography of the County is gently rolling to hilly, with several pronounced ridges. Elevations above sea level range from 587 feet (179 meters) in the southeast corner to 1,705 feet (520 meters).

Gaston County, North Carolina, is composed of the Inner Piedmont (1), Kings Mountain (2), and Charlotte (3) geologic belts (Figure 3.1) (North Carolina Department of Environmental Quality 2020). Gaston County sits astride the Central Piedmont suture zone (a complex tectonic boundary) that joins the Carolina terrane to the Cat Square terrane in the Inner Piedmont (Huebner et al. 2017). Huebner et al. (2017) described the Cat Square terrane as a remnant of an early Paleozoic ocean basin.

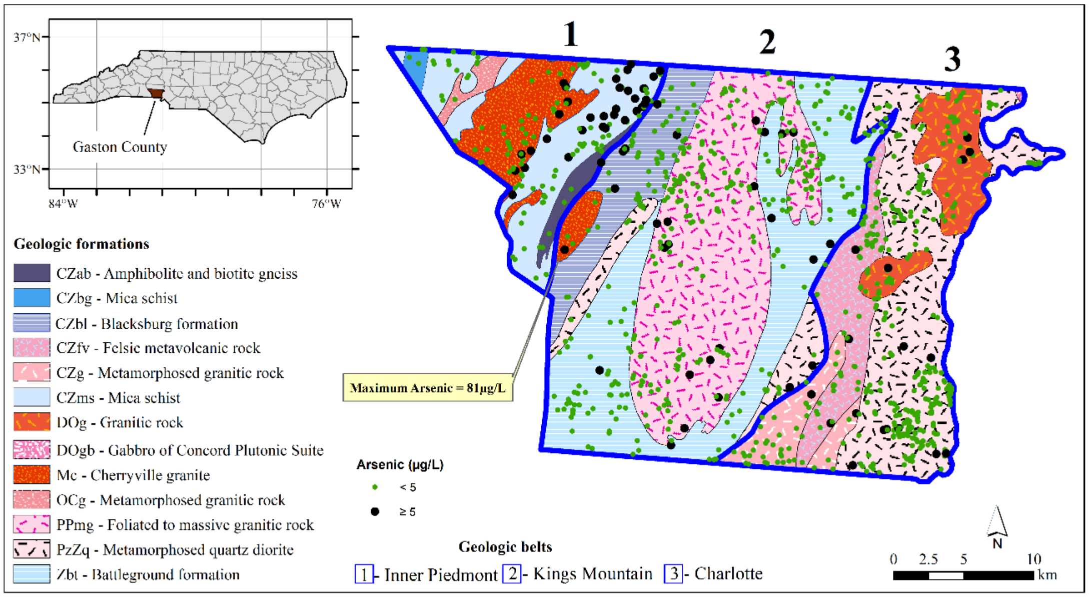
37°N
33°N
84°W
76°W
Gaston County
1
2
3
Maximum Arsenic = 81µg/L
Geologic formations
CZab - Amphibolite and biotite gneiss
CZbg - Mica schist
CZbl - Blacksburg formation
CZfv - Felsic metavolcanic rock
CZg - Metamorphosed granitic rock
CZms - Mica schist
DOg - Granitic rock
DOgb - Gabbro of Concord Plutonic Suite
Mc - Cherryville granite
OCg - Metamorphosed granitic rock
PPmg - Foliated to massive granitic rock
PzZq - Metamorphosed quartz diorite
Zbt - Battleground formation
Arsenic (µg/L)
< 5
≥ 5
Geologic belts
1 - Inner Piedmont
2 - Kings Mountain
3 - Charlotte
N
0 2.5 5 10
km

Figure 3.1: Spatial distribution of arsenic concentrations in well water samples in Gaston County, North Carolina (Geologic data source: North Carolina Department of Environmental Quality, 2020)

The geologic belts contain bedrock of varying ages and formations. A geologic formation is a fundamental unit in the classification of rocks based on similar characteristics in mineral composition, grain size, and color (Carter et al. 2002). The geologic formations in the Inner Piedmont consist of amphibolite and biotite gneiss, Cherryville granite, metamorphosed granitic rock, and mica schist (Figure 3.1). The mica schist formation consists of units abbreviated €Zs and €Zms. €Zs is a "white-mica schist" that, depending on locality, contains layers of biotite gneiss, quartz-mica schist, micaceous quartzite, and rare amphibolite. In Gaston County, much of the €Zms unit appears to be a country-rock to the Mississippian Cherryville granite (Mc in Figure 3.1) (Goldsmith et al. 1988). Following the interpretation of Goldsmith et al. (1988), the €Zms unit in this study forms a suite of mainly stratified groups of similar age and thus related source environments. The Cherryville Granite is a late- to post-metamorphic two-mica granite that is associated with elevated radon (Waldron et al. 2007; Werner et al. 2009).

The geologic formations in the Charlotte belt consist of granitic rock, metamorphosed granitic rock, metamorphosed quartz diorite, gabbro of concord plutonic suite, and felsic metavolcanic rock. The Kings Mountain belt consists of metamorphic rocks in the Battleground Formation, Blacksburg Formation, foliated to massive granitic rock, Cherryville granite, and metamorphosed quartz diorite (Goldsmith et al. 1988). The Battleground Formation consists of protoliths formed during the Late Proterozoic and later metamorphosed to form a combination of quartz-sericite schist with metavolcanic rocks, quartz-pebble metaconglomerate, and kyanite-sillimanite quartzite. The Blacksburg Formation consists of sericite schist with graphite, phyllite, amphibolite, and calc-silicate rocks formed in the late Proterozoic-Cambrian.

3.3 Material and Methods

3.3.1 Arsenic Concentration in Well Water

Arsenic data for private wells was obtained from the Gaston County Department of Health and Human Services (GC-DHHS) for 2011 through 2017. The data also contained information on the permit number, owner’s name, residential address, collection date, sampling point, pH, and other inorganic chemicals. We used a GIS to geocode residential addresses to determine their geographic coordinates (Owusu et al. 2017). Some of the records represent repeated sampling of the same well – e.g., when separatewater samples are taken from the kitchen sink and at the well. We therefore retained only the maximum recorded value from the location with the multiple tests to reflect potential groundwater concentration, which reduced our samples to 1082. This method has been used in similar studies to preserve the number of samples above the reporting limit (Ayotte et al. 2006; Kim et al. 2011; Gross and Low 2013; Ayotte et al. 2017; VanDerwerker et al. 2018). We also excluded 92 records because the pH values were missing, which reduced our final sample set to 990.

3.3.2 Determining Well Depth and Geologic Information

We obtained a digital copy of Gaston County’s private wells permit data from GC-DHHS to get well depth information to associate with the arsenic data (Figure 3.2). The well depth does not accurately reflect the depth of the water sample, because topography, groundwater flow and precipitation can affect the level of the water table. In this study, we relied on the well depth because the actual depth of the water sample was not available.

Out of the 990 samples, we were able to merge 509 arsenic samples to the permit data using either the permit numbers, residential address, or name to extract the well depth information (Figure 3.2). For the remaining 481 sampled wells that were not merged to the permit data due to missing data, we imputed the well depth information using an inverse distance weighting (IDW), an interpolation technique (Figure 3.2). The IDW surface was developed from 7837 well depths in the permit data.

Ethan and Xiao-Ming (2018) have suggested that the depth from the regolith to the bedrock aquifer frequently tapped by shallow wells ranges from 0 to 150 feet in Orange County, which is also in the Piedmont region. We assumed this could also be the case in Gaston County and classified the well depths into three groups; 1) shallow (≤ 150 feet), 2) moderate (151 – 300 feet), and deep (≥ 301 feet) to evaluate the differences in risk of arsenic in private wells. It was appropriate to categorize well depth into three groups (shallow, moderate, and deep) because the relationship between probability of arsenic concentration ≥ 5 µg/L and well depth may not be linear due to differences in well construction relative to the depths at which water enters the well. The well depth groups can therefore help to understand the real relationships and differences in the probability of arsenic concentration ≥ 5 µg/L considering that the characteristics of the sampled private wells are not provided in the data.

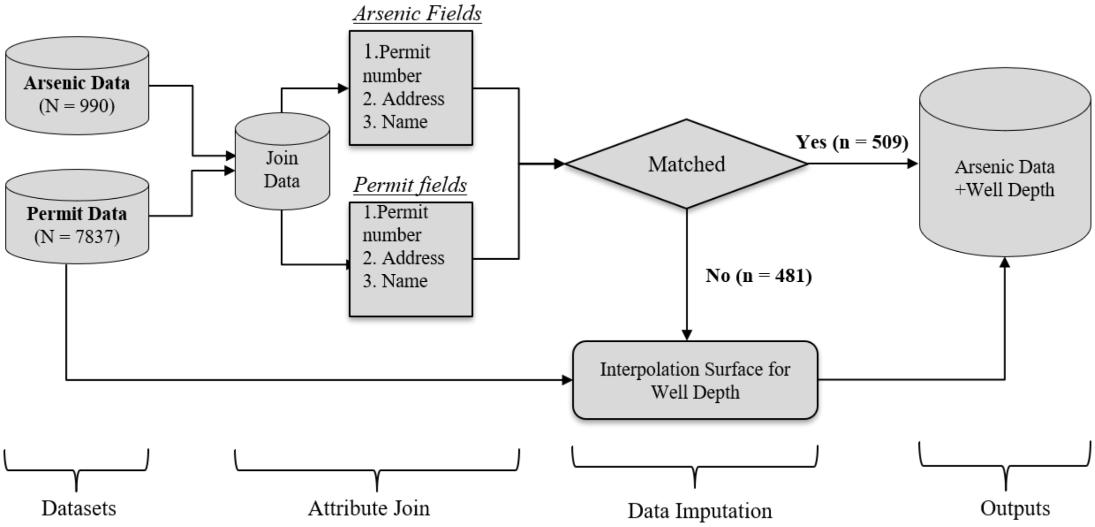

Figure 3.2: Workflow to obtain well depth for 509 samples and estimate well depth for 481 samples that have no depth information in the permit data

We obtained the geologic data for Gaston County from the North Carolina online GIS Portal. The NC Department of Environmental Quality Division of Land Resources, NC Geological Survey, and NC Center for GIS developed the digital data at a scale of 1: 250,000 miles. We spatially joined the sampled arsenic locations to the geologic data using a GIS.

3.3.3 Development of the Autologistic Regression Model

Similar to Ayotte et al. (2006), we converted the arsenic concentration to 1 if ≥ 5 µg/L and 0 if < 5 µg/L because 912 samples were marked as '< 5 µg/L' and 78 samples were reported arsenic concentrations. Because of the small number of samples with arsenic concentration ≥ 5 µg/L, we did not split the datasets into train and validation data. Instead, we used all the data in the model development to allow for a better model. We used an autologistic regression model to predict locations where the presence of arsenic concentration is ≥ 5 µg/L in private wells. The assumption for the autologistic regression is that relationships between the presence of arsenic and the explanatory factors are similar for nearby private wells than distant wells. We estimated the probability of elevated arsenic concentration at a location *i* using the autologistic function (Tsuyuki 2008).

$$p_i = \frac{1}{1 + exp\left[-\left(\beta_0 + \beta_1 x_{1,i} + ..\beta_n x_{n,i} + C(auto\ cov_i)\right)\right]} \qquad (1)$$

i is the location of the private well, $x_1 \dots x_n$ are the covariates, $\beta_0, \beta_1, \beta_n\ and\ C$ are the estimated coefficients. The introduction of the autocovariate variable in the autologistic regression penalizes the regression constant and reduces the contribution of the residuals to produce robust predictions (Griffith 2004; Dormann 2007; Fu et al. 2013; Bo et al. 2014). The autocovariate variable for a location *i* is calculated using Equation 2.

$$Auto\ cov_i = \frac{\sum_{j=1}^{k} w_{ij}\hat{P}_j}{\sum_{j=1}^{k} w_{ij}} \qquad (2)$$

The autocovariate variable ($auto\ cov$) is a weighted average of the probabilities of arsenic concentration ≥ 5 μg/L of a set of nearby private wells *j* ($j = 1 \dots k$) to the private well at *i*. The weight between private wells *i and j* is $w_{ij} = \frac{1}{d_{ij}}$, where d_{ij} is the Euclidean distance between private wells *i* and *j*, and $\hat{P}_j$ probability of arsenic concentration ≥ 5 μg/L at *j*. We determined that the minimum Euclidean distance (bandwidth) at which no private well had zero neighbors was 1976 meters and used this value (d_{ij}) in the analysis.

We used the “spatialEco” package in R/R Studio version 3.6 (Evans and Ram 2020) to implement the spatial autologistic regression model. We assessed the overall model performance by computing the receiver operating characteristic (ROC) area under curve (AUC) value. This value is a ratio of the true positive rate to the false positive rate, integrated over a range of probability thresholds, and indicates model fit (Hamel 2009). AUC values range from 0.5 to 1; where 0.5 means that the model is no better than predicting the outcome by a random chance, 0.7 is a good model; 0.8 is a robust model, and 1 is a perfect model (Hamel 2009). We also report the percentage of the correctly classified and the Chi-Square test for goodness of model fit.

3.3.4 Development of an Interpolated Probability Surface

Our model results return the probability of arsenic concentrations ≥ 5 µg/L that we mapped to reveal spatial patterns throughout Gaston County, along with the residuals using Kriging. Kriging is an interpolation method to estimate the values of a variable at unsampled locations using observations from known sites (Hengl 2009; Li and Heap 2011; 2014). The interpolation surface allows for delineating areas with a high probability of arsenic concentration ≥ 5 µg/L in well water in Gaston County. The kriging interpolation was developed with the Gstat R statistical package (Pebesma et al. 2019).

3.4 Results

3.4.1 Distribution of Arsenic Concentration

Out of the 990 arsenic measurements, a total of 912 samples contained arsenic concentrations < 5 µg/L; 78 samples were ≥ 5 µg/L. Out of 78 samples with detectable arsenic (≥ 5 µg/L), 42 samples had concentrations from 5 to 6 µg/L (Figure 3.3). The maximum reported arsenic concentration in well water was 81 µg/L in the Kings mountain geological belt (Figure 3.1). The pH in well water samples ranged from 5.1 to 9.7. The median pH in well water samples was 7.1. Sampled wells that contained arsenic concentrations ≥ 5 µg/L had an average pH of 7.3, while those at lower levels (< 5 µg/L) had a pH of 7.1. The average well depth was 242 feet with a standard deviation of 134.2. The average prediction error for the IDW was -3.1 ft and the RMSE was 125 ft.

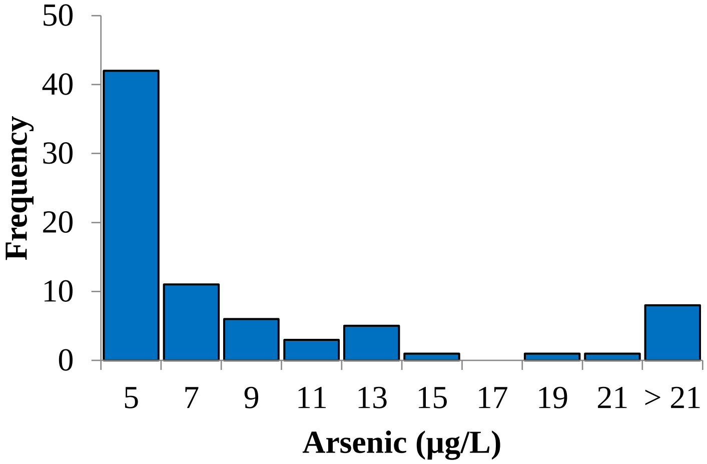

Figure 3.3: Distribution of arsenic for the 78 samples at or above detectable levels (5 µg/L) (samples marked as '< 5 µg/L' in the data had a frequency of 912- not included in the histogram)

As shown in Figure 3.1, the spatial distribution of the presence of arsenic and the geologic units in Gaston County suggest that most of the samples with arsenic concentration ≥ 5 µg/L were in the northwestern part of the county, which is an area primarily within the €Zms - Mica schist geologic unit. Specifically, within the €Zms - Mica schist unit, 28% (n = 26) of the samples in that unit exhibited arsenic concentration ≥ 5 µg/L (Table 3.1). Noteworthy, 15.1% (n = 8) of samples with arsenic concentration ≥ 5 µg/L were found in private wells identified as being located within the Mc - Cherryville Granite.

Table 3.1: Samples with arsenic concentration (≥ 5 µg/L) in geologic units in Gaston County, North Carolina

Geologic unit	Total (N)	*(n)* *(≥ 5 µg/L)*	*%* *(≥ 5 µg/L)*
€Zab - Amphibolite and biotite gneiss	7	1	14.3
€Zbg - Mica schist	4	0	0

€Zbl - Blacksburg Formation	57	5	8.8
€Zfv - Felsic metavolcanic rock	58	3	5.2
€Zg - Metamorphosed granitic rock	39	2	5.1
€Zms - Mica schist	93	26	28
DOg - Granitic rock	52	4	7.7
Mc - Cherryville Granite	53	8	15.1
O€g - Metamorphosed granitic rock	2	0	0
PPmg - Foliated to massive granitic rock	199	13	6.5
PzZq - Metamorphosed quartz diorite	279	10	3.6
Zbt - Battleground Formation	147	6	4.1

We summarized the number and percent of samples with arsenic concentration ≥ 5 µg/L for the different geologic belts in Gaston County (Table 3.2). Overall, arsenic concentrations values ≥ 5 µg/L were found in the Inner Piedmont Belt.

Table 3.2: Samples with arsenic concentration (≥ 5 µg/L) for geologic belts in Gaston County, North Carolina

		(n)	***%***
Geologic belt	**Total (N)**	(≥ 5 µg/L)	(≥ 5 µg/L)
Charlotte	409	17	4.2
Inner Piedmont	154	34	22.1
Kings Mountain	427	27	6.3

We also examined the number and percent of samples with ≥ 5 µg/L arsenic concentration by different well depth group (Table 3.3)

Table 3.3: Samples with arsenic concentration (≥ 5 µg/L) for different well depths (data includes known and simulated well depths) in Gaston County, North Carolina

		(n)	***%***
Depth (feet)	**Total (N)**	**(≥ 5 µg/L)**	**(≥ 5 µg/L)**
≤ 150	94	7	7.5
151 to 300	629	41	6.5
≥ 301	267	30	11.2

3.4.2 Model Results for Arsenic Concentration ≥ 5 µg/L

The results of the autologistic regression model adjusted for confounding factors suggests that the ЄZms - Mica schist and pH are associated with the presence of arsenic (≥ 5 µg/L) in well water (Table 3.4). The presence of arsenic ≥ 5 µg/L is significantly associated with private wells located in ЄZms - Mica schist formation, (OR = 2.99, with 95% confidence interval: (1.37 - 6.52). When adjusted for potential confounding variables, the odds of arsenic > 5 µg/L in wells on ЄZms - Mica schist was 2.99 times that of other wells. We found that one unit increase in pH in well water, the log odds of arsenic concentrations ≥ 5 µg/L increased by 0.75, when adjusted for other confounding factors. An OR= 2.11 with 95% CI: (1.31 – 3.38), indicated that arsenic concentration significantly increased with pH levels. The positive autocovariate coefficient (C= 2.80) characterizes the inherent spatial effect in the model residuals. A positive value indicates that the spatially nearby locations have similar risk patterns of arsenic concentrations being greater than 0.5. This residual spatial autocorrelation term (autocovariate) in the spatial autologistic regression helps in more accurate standard error estimates for regression coefficients.

Table 3.4: Results of significant ($p < 0.05$) variables in the spatial autologistic regression model. Positive coefficient suggests an increased probability of arsenic ≥ 5 µg/L

Variable	**Coefficient (β)**	**Odds Ratio (OR)**	**95% CI of OR**
ЄZms - Mica schist	1.09	2.99	1.37 - 6.52
pH	0.75	2.11	1.31 - 3.38
Autocovariate (C)	2.80	16.48	4.58 – 59.15

The model correctly classified 90.1% of the arsenic concentrations ≥ 5 µg/L (sensitivity) and 93.1% of the arsenic concentration < 5 µg/L (specificity). Overall, our model classification accuracy was 93.0%. The chi-square goodness of fit was significant ($p < 0.05$), which indicates that the model was better than a null model. The model AUC was 0.8, which indicates it is reliable 80% of the time in predicting the presence of arsenic concentrations ≥ 5 µg/L across Gaston County, North Carolina.

3.4.3 Spatial Autocorrelation (Autocovariate) of Arsenic Concentration ≥ 5 µg/L

The spatial distribution of the autocovariate variable represents the residual spatial autocorrelation in the autologistic regression model. The values indicate the strength of the correlation between the wells with arsenic concentrations ≥ 5 µg/L as a function of the distance separating the samples. These values range from -1 to 1, with 1 indicating areas with strong positive autocorrelation (spatial clustering), -1 indicating areas with strong negative autocorrelation, and 0 indicating a random spatial pattern with no spatial autocorrelation. As shown Figure 3.4, areas with negative values can be observed in the central part of the county (dispersion of arsenic concentrations ≥ 5 µg/L), and a large proportion of the county with zero values (random spatial pattern). We observed areas with positive spatial autocorrelation ≥ 0.41 in the northwest, northeast, and southeast areas in the county indicating samples with arsenic concentrations ≥ 5 µg/L are near each other. Having many samples with arsenic concentrations ≥ 5 µg/L near each other in the northwest, northeast, and southeast areas in the county may suggest a possible common arsenic occurrence is within the area. These areas may have a poor groundwater quality compared to other parts of the county with spatial autocorrelation <0.41. The areas with spatial autocorrelation ≥ 0.41 are consistent with the pattern in Figure 3.3 showing locations with arsenic concentrations ≥ 5 µg/L particularly in the northwest of the county.

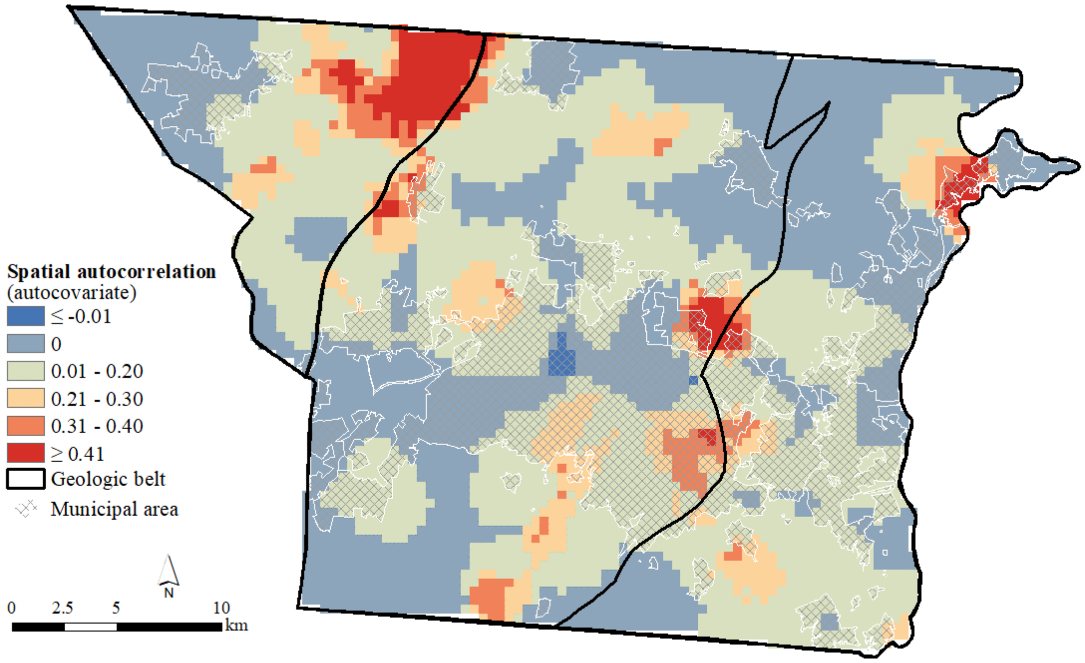

Figure 3.4: Distribution of the spatial autocorrelation (autocovariate variable)

3.4.4 Spatial Probability of Arsenic Concentration ≥ 5 µg/L

Using the model, we generated a kriging map of the probability of arsenic concentrations ≥ 5 µg/L (Figure 3.5). A probability higher than 0.5 indicated that well water was predicted to have arsenic concentration ≥ 5 µg/L, considering the combined effects of geology, pH, and well depth. Although the map shows that most places in the county have a low likelihood of arsenic concentration ≥ 5 µg/L, we can observe a high probability (> 0.5) in the northwest section of the county (Figure 3.5). This area covers about 8.4 km^2, and our model predicts that wells contained within the area have a high chance of containing arsenic concentration ≥ 5 µg/L.

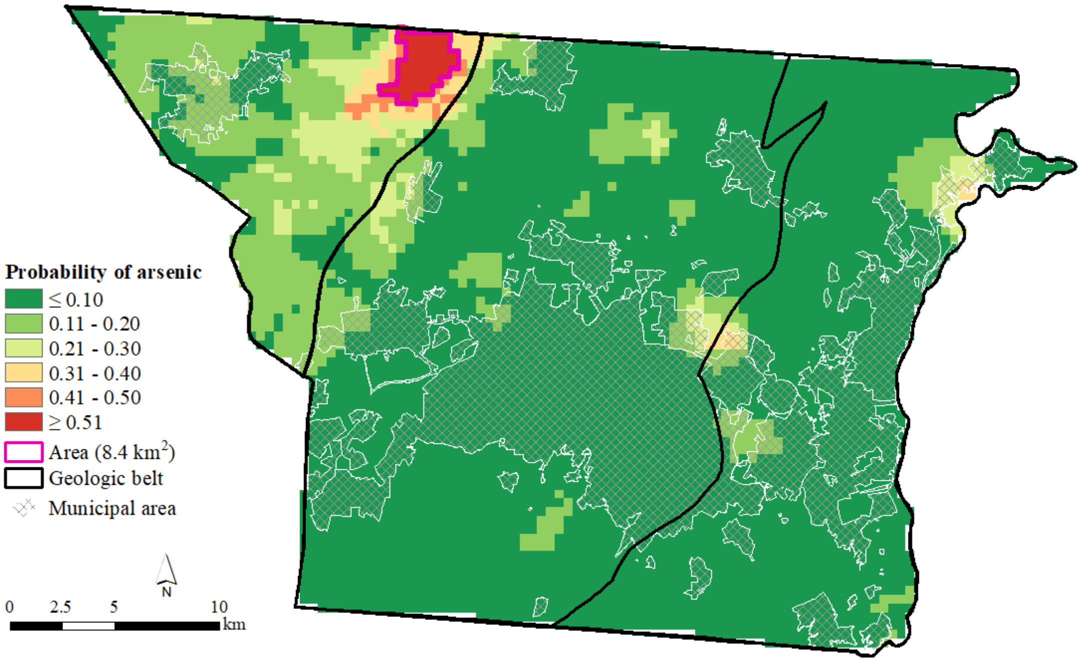

Figure 3.5: Spatial distribution of the probability of arsenic concentration ≥ 5 µg/L in private wells – model results from autologistic regression.

3.5 Discussion

Our results suggest the presence of arsenic ≥ 5 µg/L concentration in well water is related to the geologic formation and pH. We found 26.5% of the sampled wells in the Mica schist (ЄZms) formation contained arsenic concentrations ≥ 5 µg/L. This high percentage of samples with arsenic concentration ≥ 5 µg/L supports our results of the spatial autologistic regression model; private wells located in Mica schist (ЄZms) were predicted to have a threefold likelihood of having arsenic concentrations ≥ 5 µg/L after controlling for other confounding factors. Mica schist (ЄZms) formation consists of metamorphic rocks including quartz schist, micaceous quartzite, phyllite, and calc-silicate rock (Goldsmith et al. 1988; North Carolina Department of Environmental Quality 2020).

Previous studies have identified high arsenic levels in these rocks with similar assemblages of silicate rock-forming minerals (Smedley and Kinniburgh 2002; Garelick et al. 2009). The Mica schist (€Zms) formation is also part of the Inner Piedmont belt of North Carolina, a region that has been found to contain elevated arsenic concentrations (≥ 10 µg/L) in groundwater supplies due to geologic sources (Pippin 2005; Reid et al. 2005; Harden et al. 2009; Chapman et al. 2013). Our study corroborates these findings.

The 8.4 km2 area with a probability ≥ 0.5 for the presence of arsenic concentration ≥ 5 µg/L (Figure 3.5), coincides with the mica schist (€Zms) formation. Further, we observed a positive spatial autocorrelation ≥ 0.41 in the northwest (Figure 3.4) to support evidence of a possible common arsenic source related to the geology in the area. From the GIS permit database, we found that there were 75 private wells within that area, and 12 were sampled during this study period. Out of the 12 sampled private wells in the area, 75% (n=9) contained arsenic concentration ≥ 5 µg/L. The average arsenic concentration for the 9 sampled private wells was 16 µg/L. Given that lifetime exposure to even lower levels of arsenic concentration can be detrimental to human health (Mahram et al. 2013; Bräuner et al. 2014; James et al. 2015; Bloom et al. 2016; Almberg et al. 2017), all well users should be encouraged to monitor well water quality.

We found evidence that sampled wells with arsenic concentration ≥ 5 µg/L in the water had an average pH of 7.3, which may indicate slightly alkaline conditions that could increase arsenic mobilization in well water. The potential for arsenic mobilization to occur as a result of ion exchange-related increases in pH could be due to interactions between geologic minerals and aquifer waters (Ayotte et al. 2003; Ayotte et al. 2006). The pH values greater than 7.2 have closely been related to high arsenic concentrations in groundwater aquifers in the Piedmont of North Carolina (Chapman et al. 2013). Our findings corroborate these studies.

Our model results indicate no statistically significant relationship between the presence of arsenic and well depth after adjusting for other confounding factors. Similarly, we found a depth-arsenic relationship L when using data for the 509 samples with known well depth, however this relationship was statistically not significant. Previous studies have found an association between deeper wells and elevated arsenic levels (Sun 2004; Focazio et al. 2006; Kim et al. 2011; Chapman et al. 2013). Yet, our model results did not corroborate findings from these studies. We found 10.6% of sampled wells with depth ≥ 301 feet had arsenic concentration ≥ 5 µg/L. Subsequently, 7 out of 12 sampled wells in the northwestern area with an estimated 50% chance of having arsenic concentration ≥ 5 µg/L had well depth ≥ 301 feet. Given the small sample size in the most affected area, we recommend that future studies obtain more samples to determine whether there could be a relationship between arsenic concentration ≥ 5 µg/L and deeper wells.

We used publicly available data in the analysis. Thus, our approach can be applied to other areas where geologic data is available and with existing data on private wells' water quality. Also, we used a spatial autologistic regression model rather than the commonly used ordinary logistic regression model because our dependent and predictor variables were inherently spatial. The spatial autologistic regression model was used because it adjusts for spatial autocorrelation in prediction residuals due to spatial effects, which is not rectified in the non-spatial ordinary logistic regression (Tsuyuki 2008; Bo et al. 2014). A limitation of our study is that we imputed well depth information for 481 sampled wells from an IDW interpolated surface of all wells in the county. Interpolated values may not reflect the actual well depth, but we selected this approach because excluding these samples would have reduced our sample size by 49 percent. This would have affected the model statistical power and reduced our ability to find spatial patterns of the probability of arsenic concentration ≥ 5 μg/L. If we ignored the wells with interpolated depth, we would have in fact removed 31 samples with arsenic concentration ≥ 5 μg/L. Also, no significant relationship would have been found between the probability of arsenic concentration ≥ 5 μg/L and the mica schist layer using the 509 samples with known well depth. Another weakness of this study is that the number of samples that had arsenic concentration ≥ 5 μg/L was only 78. As a result, we did not split the data into training and testing during model development. We recommend additional sampling of arsenic data in the future to validate our results (Rogerson et al. 2004, Delmelle 2009).

Further, our model could be improved by the addition of other variables, including distance to potential anthropogenic arsenic sources (e.g., coal ash, landfills), geochemical, and hydrogeological conditions. Also, a more detailed geologic map such as that produced by Goldsmith et al. (1988), could allow for incorporating finer geologic information and improve the model. However, this map could not be used in this study because it was unavailable in a GIS usable format.

3.6 Conclusions

Out of 990 sampled wells, 78 contained arsenic concentration ≥ 5 μg/L, and the highest reported level was 81 μg/L. The average pH of well water in all the samples was 7.1 and ranged between 5.1 to 9.7. However, private wells with arsenic concentration ≥ 5 µg/L had an average pH of 7.3. The pH value of well water was positively associated with an increased probability of arsenic concentration ≥ 5 µg/L after controlling for confounding factors. Furthermore, the presence of arsenic ≥ 5 µg/L in well water was primarily related to private wells located in the Mica schist (€Zms) formation after controlling for other confounding factors.

The model results can be used to explain factors that are influencing arsenic occurrence at and above detectable levels. For example, the model results were utilized to investigate "where are the risk areas of arsenic at or above detectable levels?" To answer this question, kriging was used to estimate probabilities of arsenic at and above detectable levels at unsampled locations across Gaston County. From the kriging map, we identified an area in the northwestern part of Gaston County has a 50% chance of having arsenic at and above the detection limit. Further, we found a positive spatial autocorrelation ≥ 0.41 suggesting a spatial clustering of samples with arsenic concentration ≥ 5 µg/L in the northwest, northeast and southeast parts of the county which may suggest a possible common contamination source within these areas.

Our analysis further reveals that, the northwest area with spatial clustering arsenic concentration ≥ 5 µg/L and with a 50% chance of reporting elevated levels of arsenic coincided with the Mica schist (€Zms) formation. Our maps offer two relevant practical use cases - 1) private wells in the "hot spot" area can be targeted for interventions, and 2) the map can be shared with the community so well owners can take action to reduce their risk of drinking unsafe water. The model results improve our ability to predict the presence of arsenic because the area we identified as a hotspot coincide with the Mica schist and 9 out of the 12 samples in the area were at and above 5 µg/L. The model results provide evidence to warrant additional testing of wells for arsenic across Gaston County (Delmelle and Goovaerts 2019).

CHAPTER 4: PREDICTING COLIFORM PRESENCE IN PRIVATE WELLS AS A FUNCTION OF WELL CHARACTERISTICS, PARCEL SIZE AND LEACHFIELD SOIL RATING [3]

Abstract

Public water systems must test for pathogens in the water regularly, yet no testing or disinfection is required for private wells. The purpose of this study is to identify whether well age, type of well, well depth, parcel size, and soil ratings for a leachfield can predict the probability of detecting coliform bacteria in private well water using a multivariate logistic regression model. Well water samples from 1163 private wells were analyzed for the presence of coliform bacteria between October 2017 and October 2019 across Gaston County, North Carolina. The maximum well age in the data was 30 years and this reflects when the Gaston County enforced standards on well construction. The median well age for bored and drilled wells was 24 and 19 years respectively. Bored wells were shallower (mean depth = 59.2 feet) compared to drilled wells (mean depth = 259.3 feet). We found coliform bacteria in 329 samples, 290 of which were drilled wells and 39 bored wells. Using geographic information systems (GIS), we identified two areas within a 1-kilometer search radius in the northeastern part of the county in which 60 percent of 8 samples were positive for coliform bacteria. The high positive rate within the area may suggests a possible common coliform-bacteria source. The logistic regression model results showed bored wells were 4.76 times more likely to contain bacteria compared to drilled wells. In addition, we found that the likelihood of coliform bacteria significant increased with well age (average well age in data = 19 years) suggesting that those

[3] Owusu, C., Silverman, G. S, Vinson, D., Paul, R., Delmelle, E. To be submitted to *Journal of Science of the total Environment*

constructed before well standards was enforced may be at a higher risk of coliform bacteria. We found no significant association between poorly rated soils for a leachfield, well depth, parcel size and the likelihood of having coliform in wells. In conclusion, we can generalize that bored and older wells are most vulnerable to possible pathogenic contamination. These findings and our GIS maps can be leveraged to determine areas of concern to encourage well users to take action to reduce their risk of ingesting coliform bacteria.

Keywords: *Coliform Bacteria, Contamination, GIS, Private Wells, Water*

4.1 Introduction

Water contaminated with pathogens can cause illness if ingested without appropriate treatment (Wallender et al. 2014; Beer et al. 2015; Benedict et al. 2017). Illnesses that result from consuming pathogenic contaminated water includes acute gastrointestinal illness, acute respiratory illness, and neurologic illnesses (Benedict et al. 2017). Waterborne diseases may even lead to early death (Cortese and Parashar 2009; Morgan et al. 2015). For example, Chaudhry et al. (2015) found that the presence of Hepatitis E in drinking water of pregnant women increased their risk of fulminant hepatitis, deaths and may be fatal to the unborn baby.

Between 2013 and 2014, 33.3% of the total 42 waterborne disease outbreaks in public water systems in the United States (U.S.) were associated with consuming water drawn from groundwater sources (Benedict et al. 2017). These estimates do not include data on private wells; hence the actual prevalence of waterborne illness for people using groundwater sources in the U.S. is unknown (CDC 2016). As part of measures to reduce waterborne illness, the CDC recommends that well users annually test their well water for the presence of pathogens (CDC 2009). Yet, few people test their well water for contamination (Knobeloch et al. 2013; Swistock et al. 2013; Ridpath et al. 2016).

Moreover, it is virtually impossible to monitor every possible pathogen that could be in water (Toze 1999). For this reason, coliform bacteria are commonly used to indicate the possible presence of pathogens in drinking water (Toze 1999; USEPA 2017b). Coliform bacteria are microorganisms that reside in humans and animal's intestines to facilitate the breakdown of food and they are also present in soils to help with decomposition of plant materials (Farrell-Poe et al. 2010). The presence of coliform bacteria in drinking water can indicate its contamination with human or animal waste (Gerba 2009; Cabral 2010; Pandey et al. 2014). Public water systems must regularly test for total coliform in the U.S. (USEPA 2017b). However, no testing is required for private wells which at times are found in proximity pollution from septic systems (Kaplan 2014; USEPA 2015), landfills (Charrois 2010), raw manure (Sadeghi and Arnold 2002) and animal farms (Hubbard et al. 2004; Fairbrother and Nadeau 2006).

In North Carolina, approximately 48.5 percent of homes rely on septic systems to dispose of human waste (National Environmental Services Center 2019). An estimated 30,810 septic systems are used in Gaston County, North Carolina (the focus of this study, Figure 4.1) (National Environmental Services Center 2019). The widespread use of these systems in the County makes them a substantial potential contamination source. Noteworthy, septic systems are prone to failure, and effluents may go directly into an aquifer without sufficient time remove contaminants and impurities from the wastewater in a leachfield (USEPA 2015; Schaider et al. 2016).

The purification abilities of the soil for a leachfield, including absorption and filtration, dictates the detention time for pathogen removal from the wastewater (Beal et al. 2005). Absorption and filtration are influenced by the texture, porosity, and thickness of the soil layer (United States Department of Agriculture Natural Resources Conservation Service 1999). Karathanasis et al. (2006) found that course-textures soil was not as effective as fine-textured soil in removing fecal bacteria from septic system effluent. Soil suitability for a leachfield is a critical parameter to reduce contamination (USDA Natural Resources Conservation Service 2019). Furthermore, putting many leachfields next to each other (small parcels) above the same aquifer can increase the risk of incomplete wastewater treatment resulting in contamination of private wells (Yates 1985; McQuillan 2004; Swartz et al. 2006).

Many of the wells that test positive for coliforms are bored compared to drilled wells (Conboy and Goss 2000; Olabisi et al. 2008; Maran et al. 2016). Generally, bored wells are not deep compared to drilled wells (United States Geological Survey 2018). Several studies have found correlations between the presence of coliforms and shallow wells (<100 feet) (Hossain and Sivakumar 2006; Gonzales 2008; Olabisi et al. 2008; Nwachukwu et al. 2010). Shallow wells draw groundwater from the unconsolidated material that is infiltrated quickly by polluted surface water (Mesner 2012). Moreover, older wells are more likely to have thinned and pitted casing (Mesner 2012). Well age has also been positively correlated with wellhead failures and may present a route of entry for surface contamination (Sarkar et al. 2012). This study attempts to capitalize on all the various factors to predict the probability that coliform bacteria is present in a private well.

The purpose of this study is to identify whether well age, type of well, well depth, parcel size, and soil ratings for a leachfield can predict the probability of detecting coliform bacteria in private well water using a multivariate logistic regression model. Our paper is structured as follows: In Section 2, we describe our field data collection for water samples and how we secured GPS locations of the wells. Next, we describe the process to integrate well construction information (type of well, well depth, casing depth, and well age), soil ratings for leachfields, parcel size into our well sample data. We then introduce the multivariate logistic regression function to estimate the probability of finding coliform bacteria in well water. We present model results for data that include drilled and bored wells and data with only drilled wells. In Section 3, we report the results and discuss the consistency of the findings to existing studies. In Section 4, we recommend suggestions for future studies, and end with a conclusion on key implications of the findings to overall literature and study area.

Materials and Methodology

4.2 Well Sampling and Testing

Student teams from the University of North Carolina at Charlotte (UNC Charlotte) visited every permitted private well in Gaston County, North Carolina (Figure 4.1), from October 2017 to October 2019. The household's information was retrieved from the Geographic Information Systems (GIS) database of all private wells in the county (see Owusu et al. 2017 for a thorough review on the GIS database). A total of 1,302 participants consented to have their well water collected for this study. Using GIS, we would summarize the total samples and percentage of positive coliform samples into a one-kilometer grid with a bivariate map. The only purpose for this map is to explore the extent of positive coliform samples across Gaston County.

The student teams also collected geographic coordinates at the well using Mesa ^{TM} handheld GPSs (typical accuracy 2 – 5 meters)[4]. The Gaston County Department of Health and Human Services, Division of Environmental Health analyzed the water samples for the presence of coliforms within 24 hours from the time of collection using the USEPA Colisure method (USEPA 2017a).

[4] https://www.junipersys.com/Products/Mesa-Rugged-Tablet

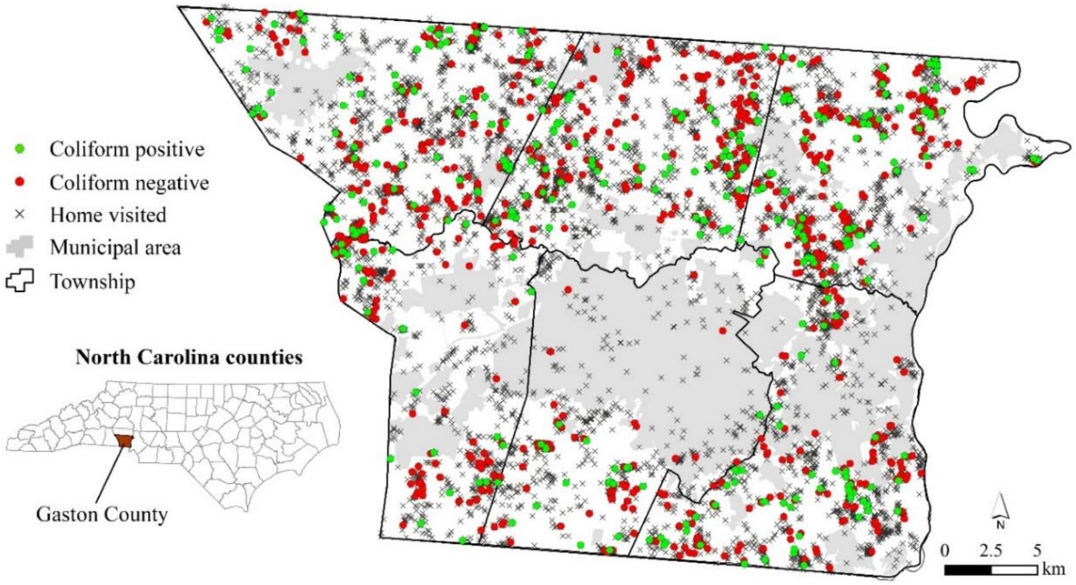

Figure 4.1: Location of positive and negative coliform samples and the homes visited from October 2017 – October 2019 in Gaston County, North Carolina

4.2.1 Incorporating Explanatory Variables into the Sampled Data for Analysis

We integrated well construction information on the type of well, well depth, and grout date into our sampled data. The well age was estimated by subtracting the grout date from the sampling date. Further, we obtained parcel data from the Gaston County Department of Planning and Development Service and extracting parcel size (acres) at each sample location. Data on 2019 soil ratings for a leachfield were obtained in GIS from the USDA Natural Resources Conservative Service for Gaston County. The soil ratings were developed from data on soil properties that affect the absorption of wastewater, the cost of construction and the replacement cost of the septic system (USDA Natural Resources Conservation Service 2019). The ratings are grouped into 1) not limited, 2) somewhat limited, and 3) very limited for use as a leachfield. In Gaston County, from the total area of 941 km^2, 320 km^2 of the soils are rated as “very limited”, 583 km^2 as “somewhat limited”, and 38 km^2 has not yet been rated for a leachfield. The county has no soil rated as “not limited” as at the time of this study. We spatially joined the sample locations to the soil suitability for leachfield data to extract the soil rating at each sample location. We excluded 139 samples because information on the type of well and age was missing.

4.3 Bivariate Statistics and Regression Modeling

We used a Chi-square statistic to test the relationships between the presence of coliform bacteria in well water and types of wells and soil ratings for a leachfield. We also examined the relationships between the presence of coliform bacteria in well water results and parcel size, well age and well depth using the Welch two-sample independent test. Next, we used multivariate logistic regression to evaluate whether the type of well, well age, well depth, parcel size, and soil ratings for a leachfield can predict the probability of finding coliform bacteria in well water. We utilized multivariate logistic regression because our response variable is binary and allows for modeling the relationships of all our input parameters to predict the probability of finding coliform bacteria in well water.

Because most of our samples were taken from drilled wells with very few samples from bored wells, we performed two statistical analysis. We first used all the data (both bored and drilled wells) in one analysis, and then also performed a second analysis with only coliform samples from drilled wells. For each analysis, we randomly selected 80% of the total data in model development and the remaining 20% for model validation purposes. We develop the multivariate logistic regression model with the Analysis of Overdispersed Data (AOD) package in R statistical software (Lesnoff et al. 2010).

4.3.1 Model Assessment

We evaluated the overall performance of the multivariate logistic regression model to predict coliform bacteria in wells using the receiver operating characteristic (ROC) area under the curve (AUC) value. This value is a ratio of the true positive rate to the false positive rate, integrated over a range of probability thresholds, and indicates model fit (Hamel 2009). AUC values range from 0.5 to 1; where 0.5 means that the model is no better than predicting the outcome by a random chance, 0.7 is a good model; 0.8 is a robust model, and 1 is a perfect model (Hamel 2009). We also reported the percentage of correctly classified in the model.

We also checked for spatial autocorrelation, which is a measure of the degree of similarity (Getis 2010) in model residuals. We tested for the presence of spatial autocorrelation using global Moran's I, available in the 'Spdep' package in R (Bivand 2009; Getis 2010). The null hypothesis for the global Moran's I states that the response variable is randomly distributed in the study area (Anselin 2019). A Moran's I value (range: -1 to 1), where negative values represent weak spatial autocorrelation, zero means no spatial autocorrelation and positive values signifies the presence of spatial autocorrelation.

4.4 Results

In the present study, the maximum well age in the data was 30 years and this reflects period when Gaston County enforced standards on well construction from 1989. Samples were obtained from 1091 drilled wells and 72 bored wells. We found well age for bored wells (median well age = 24 years) were significantly ($p < 0.05$) older than drilled wells (median age = 19 years). As shown in Figure 4.2 is the distribution of well age by the type of well.

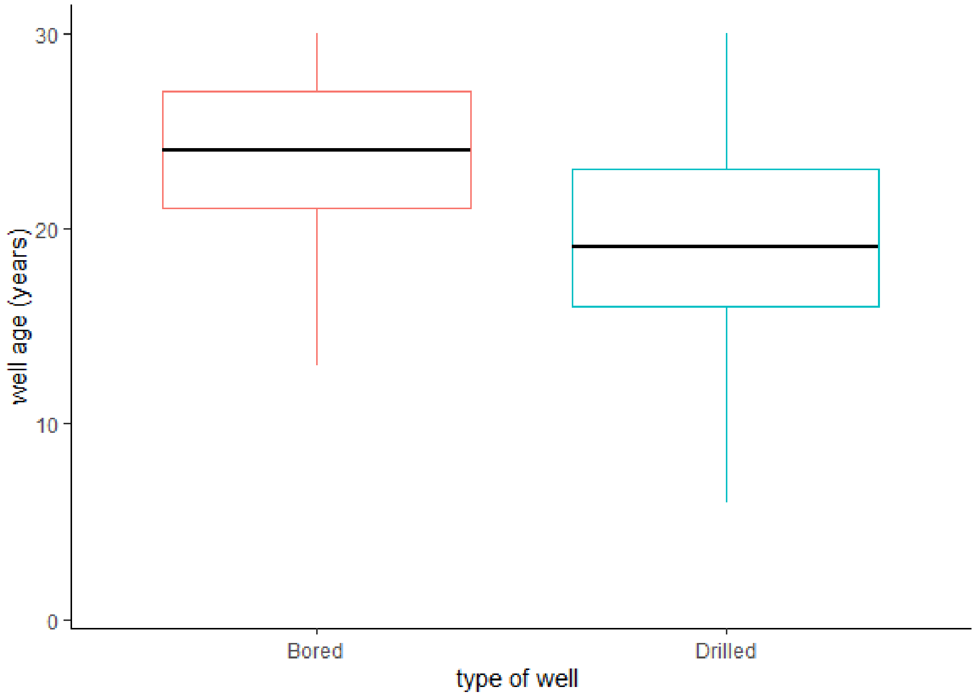

Figure 4.2: Box plot showing the distribution of well age by the type of well (median = black center line)

The median well depth for the entire data set was 230 feet. We found well depths for bored wells (median depth = 57 feet) were significantly ($p < 0.05$) shallower than drilled wells (median depth = 240 feet). As shown in Figure 4.3 is the distribution of well depth by the type of well.

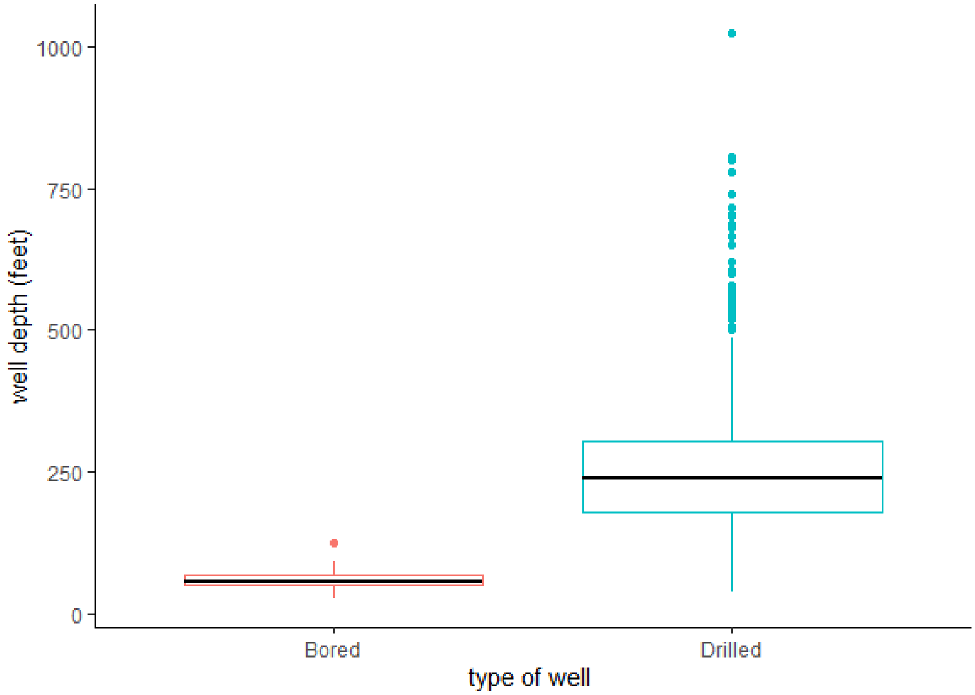

Figure 4.3: Box plot showing the distribution of well depth by the type of well (median = black center line)

Out of the total 1,163 samples, 329 (28.3%) private wells were found to be positive for coliform bacteria. The positive coliform samples were reported in 290 drilled wells and 30 bored wells. Compared to the rest of the county, a greater number of bored wells that were positive for coliform bacteria were in the northwest section of the county but appear to be evenly spaced(Figure 4.4). In contrast, drilled wells that were positive coliform were found in most townships of the county, but appear to be closer to each other particularly in the northeastern part of Gaston County.

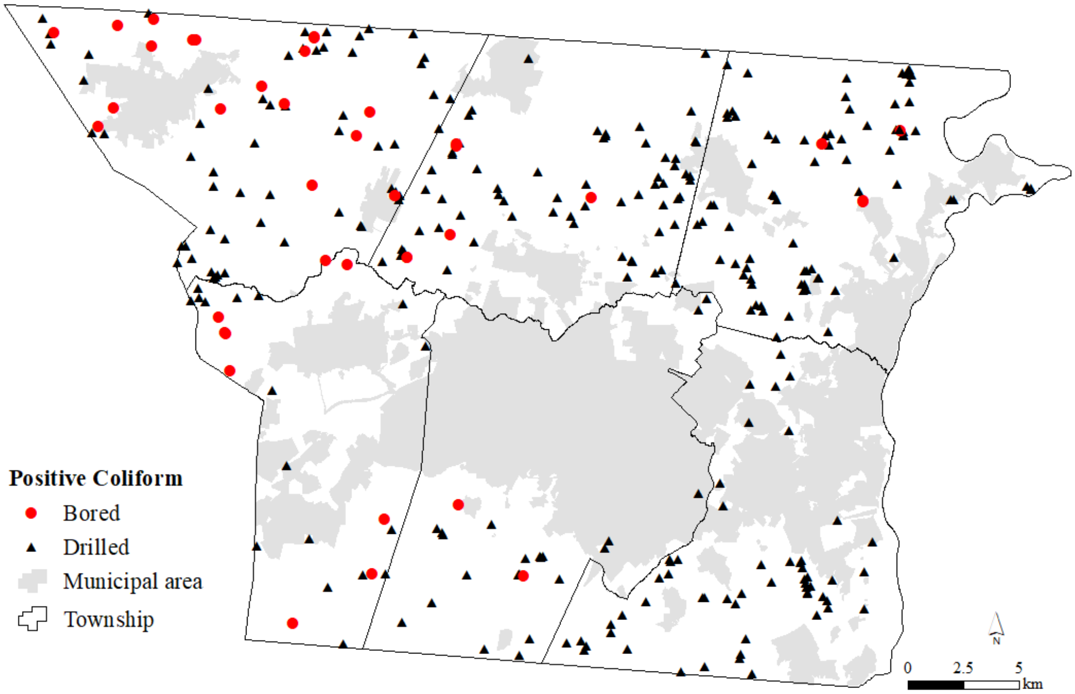

Figure 4.4:The locations of positive coliform samples by type of well

The bivariate map showing the relationship between the total number of samples and the percentage of positive samples within a 1-kilometer grid indicate that most samples were gathered in northing townships (Figure 4.5). Within the 1-kilometer area, the average number of samples were two, and in some cases one of the wells was positive for coliform bacteria. Compared to the rest of Gaston County, we observed two areas in the northeastern of the county that had >60 percent of total samples as positive coliform (Figure 4.5).

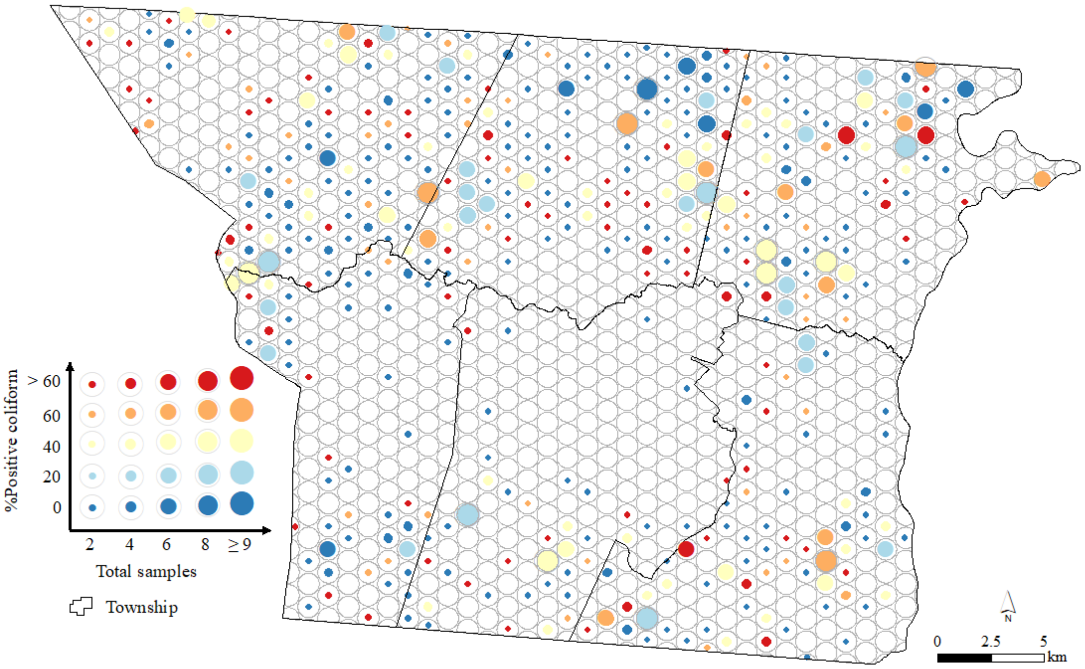

Figure 4.5: Bivariate map of percent positive coliform and total samples within a 1-kilometer grid

The characteristics of the coliform samples are shown in Table 4.1. There were significant differences between the type of wells and the presence of coliform in private wells, and wells that tested positive for coliform were likely to have been built at least 20 years ago.

Table 4.1: Characteristics of presence and absence of coliform samples in Gaston County

Characteristic	Presence (n = 329)		Absence (n = 834)	
	Count or mean	%	Count or mean	%
Well Type (count) *	329		834	
Mean well depth (feet)	251.8	--	244.9	--
Mean well age (years)*	19.7	--	18.7	--
Mean parcel size (acres)	3.3	--	3.2	--
Soil suitability for leachfield rating				
Somewhat limited (n = 958)	278	83.7	680	81.8
Very limited (n = 205)	51	16.3	154	18.2

** is p < 0.05*

As shown in Table 4.2, considering only the data for drilled wells, we found that the positive coliform samples were significantly associated with wells with deep wells (mean well depth = 277.7 ft) compared to those with no coliform (mean well depth = 252.6 ft). There was a significant difference in age for drilled wells that were positive for coliform bacteria (mean well age = 19.3 years) compared to those that had no coliform (mean well depth = 18.5).

Table 4.2: Characteristics of the sampled drilled wells with presence and absence of coliform reported in Gaston County

Characteristic	Presence (n = 290)		Absence (n = 801)	
	Count or mean	%	Count or mean	%
Mean well depth (feet)*	277.7	--	252.6	--
Mean well age (years)*	19.3	--	18.5	--
Mean parcel size (acres)	3.4	--	3.2	--
Soil suitability for leachfield rating				
Somewhat limited (n = 894)	241	83.1	653	81.5
Very limited (n = 197)	49	16.9	148	18.5

** is p < 0.05*

4.4.1 Results of the Multivariate Logistic Regression

The results of the multivariate logistic regression analyses for the two datasets; 1) data includes both drilled and bored wells (Model 1), and 2) data on only drilled wells (Model 2) are shown in Table 4.3. Overall, the type of well, and well age were significant predictors of coliform bacteria in well water, after controlling for other explanatory variables (Table 4.3). Bored wells were 3.16 times more likely to have coliform compared to drilled wells. A one-year increase in well age was expected to result in a 4%

increase in the odds of coliform bacteria being present in wells. This relationship was evident when we used all the data (both drilled and bored wells) and when drilled wells were used alone.

Table 4.3: Significant predictors of the probability of detecting coliform bacteria in well water

	Model 1		Model 2	
Predictor	Odds Ratio	95% CI	Odds Ratio	95% CI
Well Type (Drilled)				
Bored	4.73*	2.22 - 10.19		
Age (years)	1.03*	1.01 - 1.06	1.04*	1.01 - 1.06

* is $p < 0.05$; CI = confidence interval

Model 1 – data includes drilled and bored wells

Model 2 – data was only for drilled wells

4.4.2 Model Accuracy and Performance

We correctly classified 63.8% of the total data containing bored and drilled wells that had coliform bacteria (sensitivity), and 72.7% of those that had no coliform bacteria (specificity). Overall, the model classification accuracy was 70.6% for the data that include both drilled and bored wells. The performance of this model given by the AUC of 0.60 indicates that the model is reliable 60% of the time in classifying whether coliform bacteria is present or absent in a well. The AUC for the model that used data on drilled wells alone was 0.58, suggesting that the model is reliable 58% of the time in classifying whether coliform bacteria is present or absent in a sample well. However, the model did not classify any of the drilled wells to have a probability >0.5 in predicting coliform bacteria. As a result, this model should not be used in predicting coliform bacteria in wells.

4.4.3 Checking for Spatial Autocorrelation in the Model Residuals

The spatial pattern of model residuals was random, as shown by the significant ($p < 0.05$), for the data with bored and drilled wells and using only drilled wells was 0.01. Although, a positive value may signal the presence of spatial autocorrelation suggesting a need to analyze the data with a spatial model, the value of 0.01 is relatively small (values nearing 1 means higher spatial autocorrelation). The weak spatial autocorrelation in the residuals suggest that the models are not underestimating the variance of the regression coefficients to necessitate using a spatial model to account for the spatial effects (Rogerson 2019). Thus, multivariate logistic regression models were adequate to establish the relationships between the presence of coliform in well water and the predictor variables.

4.4.4 Trends in the Probability of Coliform Bacteria in Wells

We summarize the relationships between the types of wells, the ratio of casing depth to well depth and age of the well, and coliform bacteria in well water in Figures 3 to 5. Out of the 1091 drilled wells, 725 (66.5%) had a probability of ≤ 0.3 of detecting coliform bacteria well water (Figure 4.6). Out of the 72 bored wells, 40 (55.6%) had a probability of ≥ 0.5 of detecting coliform bacteria well water.

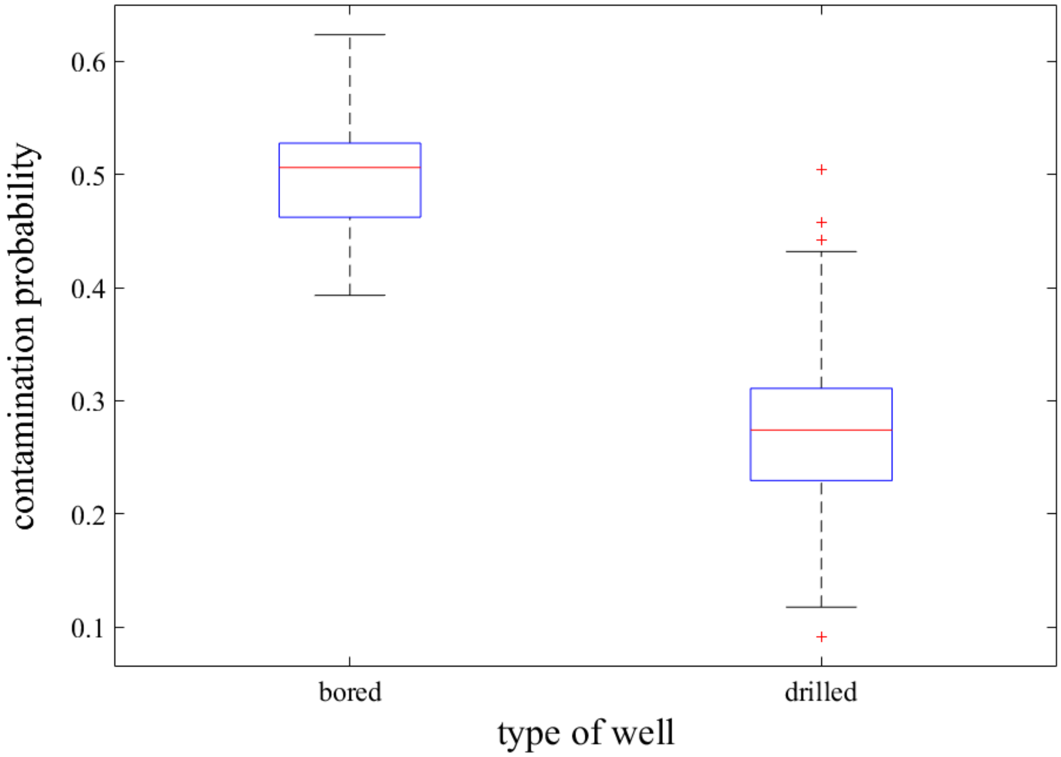

Figure 4.6: Boxplot of the probability of finding coliform bacteria in well water by types of private well using Model 1 (red central mark = median, and the bottom and top edges of the box are the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers; outliers plotted with the red '+' symbol.)

Figure 4.7 is a summary of the probability of detecting coliform bacteria in well water by well age. Overall, there was a significant positive association between the well age and the probability of detecting coliform bacteria in well water in both models. Nearly 29% of the variability in probability of detecting coliform bacteria in well water was explained by well age in the model results that used both bored and drilled wells (Figure 4.7a). In comparison, 45% of the variability in the probability of detecting coliform bacteria was explained by the age of the drilled well (using only sampled drilled wells (Figure 4.7b).

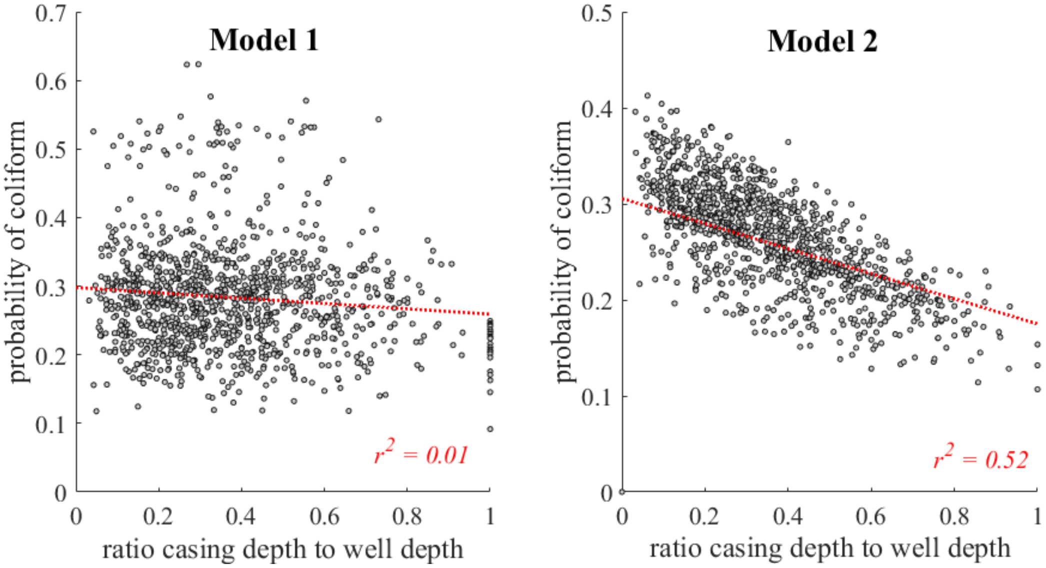

Figure 4.7: Probability of detecting coliform bacteria in well water by the ratio of casing depth to well depth, for models 1 and 2.

4.5 Discussion

Coliform bacteria was reported in 290 drilled wells and 39 bored wells. Coliform bacteria was more likely to occur in older wells. A high proportion of bored wells that were positive for coliform bacteria were found in the northwestern part of the county. Compared to bored wells, drilled wells that were positive for coliform bacteria were closer to each other. In some instances, out of more than nine samples within a 1-kilometer area in the northeastern parts of the county, 60 percent of the samples were positive for coliform bacteria. The high positive rate in such areas may suggests a possible common coliform-bacteria source.

Out of a total of 72 bored wells, 39 (54.2%) had coliform bacteria. Bored wells were 3.16 times more likely to have coliform bacteria compared to drilled wells after adjusting for all other confounding variables. Our finding corroborates other studies suggesting that bored wells are more likely to test positive for coliform than drilled wells

(Olabisi et al. 2008; Hynds et al. 2012; Maran et al. 2016). Our results are also consistent with other studies that have found that bored are at elevated likelihood of being contaminated because the wells are shallow as shown by the presence of coliform (Godfrey et al. 2006; Gonzales 2008; Olabisi et al. 2008; Nwachukwu et al. 2010; Maran et al. 2016). Our results highlight the need to evaluate the appropriateness of issuing well construction permits for bored wells.

Private wells that contained coliform were more likely to have a mean age of 20 years compared to mean age of 19 years for those with no coliforms. older. In addition, there was a positive association between well age and the probability of detecting coliform after controlling for all other confounding factors. The effect of well age on the presence of coliform may be due to failures of the wellhead, well screens, and well casings as they become older (Mesner 2012; Sarkar et al. 2012). Our results collaborate with these findings. Older wells may need to be tested often because of an increased risk of contamination.

We did not find any association between the presence of coliform bacteria in well water and parcel size and soil ratings for a leachfield. One hypothesis why parcel size may not be associated with the presence of coliform bacteria in well water for the present study could be the enforcement of a private well ordinance in Gaston County, North Carolina, since 1989. The law only permits setting a leachfield at least 100 feet from a private well (The Gaston County Board of Health 2011). Enforcement of this ordinance may have played a role in not locating septic tanks near one another, thus reducing the risk of coliform bacteria in private wells. The soil rating for a leachfield may not be the best variable to determine whether coliform bacteria were present wells. The lack of significant relationships may suggest that the presence of coliform in wells may come from runoff water. Also, because the ratings are based on more than one variable, future studies should examine the wealth of the variable used in computation of the ratings. For example, neural networks can be used to determine whether there are more complex relationships than were originally anticipated in the variables used to derive the ratings. In addition, there was no significant association between well depth and coliform presence in wells.

The major strength of our study is we combine GIS maps and multivariate logistic regression model to uncover the extent of positive rate of coliform samples across the county and identify significant predictors of coliform bacteria. For example, the northeastern part of the county where within a 1-kilometer area, out of more than 9 wells, 60% of the wells were positive for coliform bacteria indicate a possible common contamination source. From our model results we identified that well age, and bored wells are significant predictors of coliform bacteria being present which are consistent with past studies.

Despite an improved understanding of potential factors associated with the presence of coliform bacteria in the study area, additional data such as the conditions of the wellhead and well screens may account for some variation in risk, but this data was not available for inclusion in our analysis. Another limitation of our study is that we did not include measures of the proximity of sampled wells to other possible sources of biological contamination, including landfills, poultry, and dairy farms. Including more variables in the analysis in future studies may ultimately lead to finding discernable spatial patterns in coliform bacteria present in well water across the county.

Furthermore, private wells database in Gaston County does not include wells installed prior to 1989. This is because well construction permit requirement was required from 1989 after the enactment of the well ordinance (The Gaston County Board of Health 2011). Although far older wells (> 30 years) may be at a higher risk of having coliform bacteria than the wells considered in this study, the lack of data on these wells is a limitation of this study.

4.6 Conclusions

We found evidence of groundwater contamination with regards to the U.S. EPA standards for the presence of coliform bacteria in water. More than one-fourth of private wells were identified to have coliform bacteria. We identified that within a 1-kilometer search radius, there are two areas in the northeastern part of the county in which 6 out of 8 samples being positive for coliform. This may suggest a possible common coliform-bacteria source nearby. However, in this study, we could not determine the possible coliform bacteria source due to lack of available data at the scale of our analysis.

The positive rate of coliform bacteria samples was more for bored wells (54.2%) compared to drilled wells (26.6%). The multivariate logistic regression indicates that bored wells were 4.73 times more likely to have coliform bacteria compared to drilled wells. Our finding that bored wells are highly predictive of the presence of coliform bacteria in well water is consistent with past studies. Older wells were significantly related to the probability of detecting coliform in well water. Bored wells (median well age = 24 years) were significantly older than drilled wells (median well age = 19 years).

The present study shows that data analysis of the presence of coliform bacteria in combination with well age, type of well, well depth, parcel size, and soil ratings for a leachfield using multivariate logistic regression and GIS maps can provide preliminary insight on causes and extent of groundwater contamination in the county. Private well users should be mindful of the increased risk of possible pathogenic contamination when using bored wells. The appropriateness of issuing permits for bored wells may need to be considered to protect human health. Older and bored wells need to be tested often because of higher probability of coliform bacteria being present so appropriate actions can be taken to prevent consuming unsafe water.

CHAPTER 5: GENERAL DISCUSSION AND CONCLUSIONS

The goal of this dissertation was to incorporate multilevel geocoding and spatial modeling techniques to predict the risk of arsenic and coliform bacteria in wells. As a result, a multi-stage approach to geocoding was to re-engineer input addresses using probabilistic record linkage, secondary address information on past coliform and chemical test results and geocode the addresses with rooftop, parcel and street geocoding techniques. The approach improved the geocoding match rate from 38.0% to 98.9%. If the input addresses were not re-engineered, 50.9% of the total records would have been removed from the GIS database. The hypothesis that improving input addresses translates into an increase match rate is duly justified. The approach is relevant to studies concerned with utilizing address geocoding as a key research methodology to select samples for further analysis.

Further, GPS coordinates were acquired from 1075 private wells to evaluate the positional accuracy of rooftop, parcel, and street geocodes. The results indicated that there are significant differences in the positional accuracy for rooftop, parcel, and street geocoding. The results suggest when the GPS coordinates are unavailable, rooftop geocodes (mean positional accuracy = 26 meters) may offer a better representation of the location of a private well, followed by parcel geocodes (mean positional accuracy = 44 meters) before considerations are given to street geocodes (mean positional accuracy = 72 meters). The differences clearly show there is inherent uncertainty in the position of geocodes relative to the actual ground.

The findings show that positional uncertainty in geocodes may be reduced if rooftop geocoding is used, contributing to the growing concern for geographers and GIS scientists to access spatial data uncertainty. In this dissertation, 93% of GIS data on private wells were obtained from rooftop geocoding, 3.5% from parcel geocoding, and 3.1% from street geocoding – suggesting positional uncertainty was minimized for a bulk of the geocoded private wells in GIS database.

The goal of Chapter 3 was to evaluate if the geology, pH, and well depth can improve the prediction of arsenic at or above detectable levels (≥ 5 μg/L) found in private wells. A spatial autologistic regression model was developed because the arsenic distribution exhibited spatial patterns. The results indicated that the presence of arsenic was significantly associated with private wells located in ₠Zms - mica schist formation. The mica schist formation consists of assemblages of silicate rock-forming minerals known to contain high levels of arsenic. Further, there was a significant positive association between pH and the arsenic presence in wells, which may suggest possible desorption of arsenic from metal or clayey minerals. This is the first study of its kind to apply spatial autologistic regression to predict arsenic presence. The model results can be used to explain questions related to “why,” “where,” and “what” factors are influencing arsenic occurrence at or above detectable levels. For example, the model results were utilized to investigate “where are the risk areas of arsenic detectable levels?” To answer this question, kriging was used to estimate probabilities of arsenic at or above detectable levels across Gaston County. The kriging map identified that an area (8.4 km2) in the northwestern section of the county has 50% chance of having arsenic at or above the detection limit. The map offers two relevant practical use cases - 1) private wells in the “hot spot” area can be targeted for interventions, and 2) the map can be shared with the community so well owners can take action to reduce their risk of drinking unsafe water.

In addition, compared to the RMSE from the indicator kriging results, the spatial autologistic regression produced a smaller RMSE (0.252. This may be due to the addition of explanatory variables in the spatial autologistic regression which helps reduce the uncertainty in model prediction when few samples with arsenic concentration ≥ 5 µg/L are found in an area. In such instances, combine effect of the explanatory variable can help attenuate errors in the prediction to produce robust estimates.

Chapter 4 was to identify whether well characteristics, parcel size, and soil ratings for a leachfield can predict the probability of detecting coliform bacteria in wells using multivariate logistic regression. The results indicated that bored and older wells are more likely to have coliform bacteria. The specific coliform bacteria source was not determined from this study due to lack of available data on pollution sources. The models in Chapter 4 can be used to explain “why” and “what” factors are influencing coliform bacteria presence in wells.

The geocoding approach and results obtained in Chapter 2 were crucial for Chapters 3 and 4. For instance, the geocoding approach in Chapter 2 was adopted for Chapter 3 to obtain geographic coordinates of private wells in the arsenic data, yielding a match rate of 100% (n = 990). These coordinates were from rooftop geocodes (n = 956), and parcel geocodes (n = 34), suggesting geographic coordinates were adequate to perform further spatial analysis in Chapter 3 since rooftop geocodes are good representation of private wells in GIS. Subsequently, well depth information was obtained for modeling arsenic presence by joining records in arsenic data, and the GIS database developed in Chapter 2. Furthermore, in this dissertation, student teams were sent to the geocoded addresses in the GIS database of private wells (results of chapter 2) to administer free water sampling for coliform bacteria. Samples gathered for that exercise translated to the data used in the coliform analysis in Chapter 4. The well characteristics parameters used in developing the model for coliform bacteria presence in Chapter 4 were obtained from the GIS database of private wells. In summary, it is imperative to create an accurate GIS database of private wells to facilitate groundwater quality analysis.

Analyzing both arsenic and coliform bacteria risk in private wells allowed for shedding light on the different pathways for well water to become unsafe for drinking. Detectable levels of arsenic were associated with natural sources of arsenic in the geology and pH, and the presence of coliform bacteria was associated with well characteristics (well age and depth). The results capture the complexities of how environmental contaminants can enter wells. Yet, private wells are not regulated anywhere in the U.S. This dissertation may facilitate discourse on why regulations on private wells might be useful.

Further, the results show the appropriateness for issuing permits for bored wells needs to be evaluated. This is because bored wells were more than fourfold at-risk of coliform bacteria compared to drilled wells, and this is consistent with past studies. It can therefore be assumed that people obtaining water from bored wells would possibly be exposed to contamination sometime during their lifetime use of the well. It may also be appropriate to consider the incorporation of well type in real estate disclosures, providing new homeowners with information on the risks and mitigation strategies associated with bored wells.

Moreover, an area in the northwestern section (8.4 km^2) of the county was found to have an average arsenic concentration of 16 µg/L. High concentrations in this region are significantly associated with geology. Given that it is virtually impossible to remove the bedrock to reduce the burden, perhaps, expanding a neighboring municipal water system or providing open taps in neighboring communities could decrease dependence on wells in this area and help provide safe drinking water for the residents.

The success of the analysis in this dissertation was due to the development of a GIS database. The results of this database would be shared with public health officials and their ability to frequently update the database with test results may eventually lead to a reliable surveillance data of private wells, which is critical to monitoring water quality. In order to achieve this, the workforce may need to be trained on GIS operations in geocoding new permits and update attribute information with test results.

Most important, educating well owners on potential contamination sources and risk of exposure may have an impact on their willingness to test their wells frequently. Based on the results of this dissertation, targeted educational programs are needed immediately for the well owners in the northwestern area with a high probability of arsenic with the Mica schist formation. Well owners in the area could be encouraged to install treatment systems to remove arsenic from their drinking water. Also, all bored well users should be encouraged to take action to reduce their risk of exposure to possible pathogenic contamination.

Although results are useful to improving water quality in Gaston County, and may be adopted to other areas, this study also had limitations that should be address in future studies. To begin, a holistic approach to examine environmental hazard is to determine the hazard-exposure-dose-response process (National Research Council 1991). Specifically, the source of contamination (hazard), health outcome (exposure), the quantity of the contaminant in the human body (dose), and time for symptoms to show (response) are modeled as a series of interconnected processes. However, the present study did not include health information to account for exposure, dose, and response. This could be a major limitation on the validity of the results considering that point of a sample of water was at the well, and owners may use treatment systems to filter the contaminants. To validate model results for arsenic, health data on residents in the county should be examined in relation to exposure to arsenic and coliform bacteria in private wells. Also, the spatial autologistic regression used in modeling arsenic at or above detectable levels did not allow for incorporating temporal information in the analysis. Future studies should examine whether the model could be extended to account for temporal variation in arsenic samples. Furthermore, proximity measures to other possible sources of arsenic and coliform bacteria not examined in this dissertation should be considered in future studies to improve the models. Subsequently, the accuracy of the geological map may affect the results of this study and future studies should examine this issue in detail. Future studies in the area can also consider other environmental hazards (e.g. benzene, lead, sulphur) and take full advantage of available data on underground storage tanks, locations of facilities on the national toxic release inventory (TRI) to enrich the understanding of groundwater contamination in the county. More work is necessary from a 'qualitative' perspective to better understand the barriers that exist to access clean water in private wells.

REFERENCES

Chapter 1

Abokifa, A. A., Katz, L., & Sela, L. (2020). Spatiotemporal trends of recovery from lead contamination in Flint, MI as revealed by crowdsourced water sampling. Water Research, 171, 115442.

Agency for Toxic Substances and Disease Registry (ATSDR) (2020). The ATSDR 2019 Substance Priority List. Retrieved from: https://www.atsdr.cdc.gov/spl/index.html

Ali, M., Emch, M., Donnay, J. P., Yunus, M., & Sack, R. B. (2002). The spatial epidemiology of cholera in an endemic area of Bangladesh. *Social science & medicine*, *55*(6), 1015-1024.

Almberg, K. S., Turyk, M. E., Jones, R. M., Rankin, K., Freels, S., Graber, J. M., & Stayner, L. T. (2017). Arsenic in drinking water and adverse birth outcomes in Ohio. *Environmental Research, 157*, 52-59. doi:10.1016/j.envres.2017.05.010

Antunes, I. M., & Albuquerque, M. T. (2013). Using indicator kriging for the evaluation of arsenic potential contamination in an abandoned mining area (Portugal). *Science of the Total Environment, 442*, 545-552. doi:10.1016/j.scitotenv.2012.10.010

Ayotte, J. D., Medalie, L., Qi, S. L., Backer, L. C., & Nolan, B. T. (2017). Estimating the high-arsenic domestic-well population in the conterminous United States. *Environmental Science & Technology, 51*(21), 12443-12454.

Ayotte, J. D., Montgomery, D. L., Flanagan, S. M., & Robinson, K. W. (2003). Arsenic in groundwater in eastern New England: occurrence, controls, and human health implications. *Environmental Science & Technology*, *37*(10), 2075-2083.

Ayotte, J. D., Nolan, B. T., Nuckols, J. R., Cantor, K. P., Robinson, G. R., Baris, D., . . . Silverman, D. T. (2006). Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. *Environmental Science & Technology, 40*(11), 3578-3585.

Beal, C., Gardner, E., & Menzies, N. (2005). Process, performance, and pollution potential: A review of septic tank–soil absorption systems. *Soil Research, 43*(7), 781-802.

Beer, K. D., Gargano, J. W., Roberts, V. A., Hill, V. R., Garrison, L. E., Kutty, P. K., . . . Yoder, J. S. (2015). Surveillance for waterborne disease outbreaks associated with drinking water—United States, 2011–2012. *MMWR. Morbidity and mortality weekly report, 64*(31), 842.

Bellander, T., Berglind, N., Gustavsson, P., Jonson, T., Nyberg, F., Pershagen. G., Järup, L. (2001). Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. *Environmental health perspectives,* 109:633-639.

Benbrahim-Tallaa, L., & Waalkes, M. P. (2008). Inorganic arsenic and human prostate cancer. *Environmental health perspectives, 116*(2), 158-164.

Benedict, K. M., Reses, H., Vigar, M., Roth, D. M., Roberts, V. A., Mattioli, M., . . . Fullerton, K. E. (2017). Surveillance for waterborne disease outbreaks associated with drinking water—United States, 2013–2014. *MMWR. 66*(44), 1216.

Bentley, J. P., Ford, J. B., Taylor, L. K., Irvine, K. A., & Roberts, C. L. (2012). Investigating linkage rates among probabilistically linked birth and hospitalization records. *BMC medical research methodology, 12*(1), 149.

Blackburn, B. G., Craun, G. F., Yoder, J. S., Hill, V., Calderon, R. L., Chen, N., ... & Beach, M. J. (2004). Surveillance for waterborne-disease outbreaks associated with drinking water—United States, 2001–2002. *MMWR surveill summ,* 53(8), 23-45.

Bloom, M. S., Neamtiu, I. A., Surdu, S., Pop, C., Anastasiu, D., Appleton, A. A., . . . Gurzau, E. S. (2016). Low level arsenic contaminated water consumption and birth outcomes in Romania—An exploratory study. *Reproductive Toxicology, 59,* 8-16.

Bo, Y. C., Song, C., Wang, J. F., & Li, X. W. (2014). Using an autologistic regression model to identify spatial risk factors and spatial risk patterns of hand, foot and mouth disease (HFMD) in Mainland China. *BMC Public Health, 14*(1), 358.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., & Freudenheim, J. L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology, 14*(4), 408-412.

Bräuner, E. V., Nordsborg, R. B., Andersen, Z. J., Tjønneland, A., Loft, S., & Raaschou-Nielsen, O. (2014). Long-term exposure to low-level arsenic in drinking water and diabetes incidence: a prospective study of the diet, cancer and health cohort. *Environmental health perspectives*, *122*(10), 1059-1065.

Briggs, D. J. (2000). *Environmental health hazard mapping for Africa*: WHO-AFRO Harare.

Cabral, J. P. (2010). Water microbiology. Bacterial pathogens and water. *International journal of environmental research and public health, 7*(10), 3657-3703.

Cayo, M. R., & Talbot, T. O. (2003). Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics, 2*(1), 10.

Centers for Disease Control and Prevention (CDC). (2015). Arsenic and Drinking Water from Private Wells. Retrieved from: https://www.cdc.gov/healthywater/drinking/private/wells/disease/arsenic.html

Centers for Disease Control and Prevention (CDC). (2019). Safer Well Water through Stronger Public Health Programs. Retrieved from https://blogs.cdc.gov/yourhealthyourenvironment/2019/11/25/safer-well-water-through-stronger-public-health-programs/

Chapman, M. J., Cravotta, C., Szabo, Z., & Lindsey, B. D. (2013). *Naturally Occurring Contaminants in the Piedmont and Blue Ridge Cystalline-rock Aquifers and Piedmont Early Mesozoic Basin Siliciclastic-rock Aquifers, Eastern United States, 1994-2008*: US Department of the Interior, US Geological Survey.

Charrois, J. W. (2010). Private drinking water supplies: challenges for public health. *Cmaj, 182*(10), 1061-1064.

Chaudhry, S. A., Verma, N., & Koren, G. (2015). Hepatitis E infection during pregnancy. *Canadian Family Physician,* 61(7), 607-608.

Chou, Y.-H. (1995). *Automatic bus routing and passenger geocoding with a geographic information system.* Paper presented at the Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future.

Christakos G, Serre ML (2000) Spatiotemporal analysis of environmental exposure–health effect associations. *Journal of exposure science & environmental epidemiology* 10:168-187.

Cockings, S, Dunn, C. E, Bhopal, R. S, Walker DR (2004) Users' perspectives on epidemiological, gis and point pattern approaches to analysing environment and health data. *Health & Place* 10:169-182.

Conboy, M., & Goss, M. (2000). Natural protection of groundwater against bacteria of fecal origin. *Journal of contaminant hydrology, 43*(1), 1-24.

Cortese, M. M., & Parashar, U. D. (2009). Prevention of rotavirus gastroenteritis among infants and children: recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbidity and Mortality Weekly Report: Recommendations and Reports*, *58*(2), 1-25.

Cressie, N., & Kornak, J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical science*, 436-456.

Cromley, E. K., & McLafferty, S. L. (2011). *GIS and public health*: Guilford Press.

Cutchin, M. P. (2007). The need for the "new health geography" in epidemiologic studies of environment and health. *Health & place*, *13*(3), 725-742.

Data.Gov (2020). Map of Arsenic concentrations in groundwater of the United States from 31,000 wells and springs in 49 states compiled by the United States Geological Survey (USGS). Retrieved from: https://catalog.data.gov/dataset/map-of-arsenic-concentrations-in-groundwater-of-the-united-states

Dauphiné, D. C., Smith, A. H., Yuan, Y., Balmes, J. R., Bates, M. N., & Steinmaus, C. (2013). Case-control study of arsenic in drinking water and lung cancer in California and Nevada. *International journal of environmental research and public health*, *10*(8), 3310-3324.

Delmelle, E. M., Cassell, C. H., Dony, C., Radcliff, E., Tanner, J. P., Siffel, C., & Kirby, R. S. (2013). Modeling travel impedance to medical care for children with birth defects using Geographic Information Systems. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 97(10), 673-684.

Dieter, C., Linsey, K., Caldwell, R., Harris, M., Ivahnenko, T., Lovelace, J., . . . Barber, N. (2018). Estimated use of water in the United States county-level data for 2015 (ver. 2.0, June 2018). US Geological Survey data release, 10. doi:https://doi.org/10.5066/F7TB15V5

Dormann, C. F. (2007). Assessing the validity of autologistic regression. *Ecological modelling*, *207*(2-4), 234-242.

Dummer, T. J. (2008). Health geography: supporting public health policy and planning. *Cmaj*, *178*(9), 1177-1180.

Dummer, T., Yu, Z., Nauta, L., Murimboh, J., & Parker, L. (2015). Geostatistical modelling of arsenic in drinking water wells and related toenail arsenic concentrations across Nova Scotia, Canada. *Science of the Total Environment*, *505*, 1248-1258.

Dyal, J. W. (2020). COVID-19 Among Workers in Meat and Poultry Processing Facilities―19 States, April 2020. *MMWR. Morbidity and mortality weekly report*, 69.

Elliott, P., Briggs, D., Morris, S., de Hoogh, C., Hurt, C., Jensen, T. K., ... & Jarup, L. (2001). Risk of adverse birth outcomes in populations living near landfill sites. *Bmj, 323*(7309), 363-368.

Fairbrother, J., & Nadeau, E. (2006). Escherichia coli: on-farm contamination of animals. *Rev Sci Tech, 25*(2), 555-569.

Farrell-Poe, K., Jones-McLean, L., & McLean, S. (2010). Microorganisms in Private Water Wells. Retrieved from https://repository.arizona.edu/handle/10150/156925

Flora, S. J. (2011). Arsenic-induced oxidative stress and its reversibility. *Free Radical Biology and Medicine, 51*(2), 257-281.

Fox, M. A., Nachman, K. E., Anderson, B., Lam, J., & Resnick, B. (2016). Meeting the public health challenge of protecting private wells: Proceedings and recommendations from an expert panel workshop. *Science of the Total Environment, 554-555*, 113-118. doi:https://doi.org/10.1016/j.scitotenv.2016.02.128

Fu, R., Thurman, A. L., Chu, T., Steen-Adams, M. M., & Zhu, J. (2013). On estimation and selection of autologistic regression models via penalized pseudolikelihood. *Journal of agricultural, biological, and environmental statistics, 18*(3), 429-449.

Gaus, I., Kinniburgh, D., Talbot, J., & Webster, R. (2003). Geostatistical analysis of arsenic concentration in groundwater in Bangladesh using disjunctive kriging. *Environmental Geology, 44*(8), 939-948. doi:10.1007/s00254-003-0837-7

Gerba, C. P. (2009). Indicator microorganisms. In *Environmental microbiology* (pp. 485-499): Elsevier.

Goldberg, D. W. (2011). Improving Geocoding Match Rates with Spatially-Varying Block Metrics. *Transactions in GIS, 15*(6), 829-850.

Goldberg, D. W., Ballard, M., Boyd, J. H., Mullan, N., Garfield, C., Rosman, D., . . . Semmens, J. B. (2013). An evaluation framework for comparing geocoding systems. *International Journal of Health Geographics, 12*(1), 50.

Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2007). From text to geographic coordinates: the current state of geocoding. *19*(1), 33.

Gonzales, T. R. (2008). The effects that well depth and wellhead protection have on bacterial contamination of private water wells in the Estes Park Valley, Colorado. *Journal of Environmental Health, 71*(5), 17-23.

Goodchild, M. F. (2009). Geographic information systems and science: today and tomorrow. *Annals of GIS, 15*(1), 3-9.

Goovaerts, P. (2009). AUTO-IK: a 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. *Computers & geosciences, 35*(6), 1255-1270.

Goovaerts, P. (2017a). The drinking water contamination crisis in Flint: Modeling temporal trends of lead level since returning to Detroit water system. Science of the Total Environment, 581, 66-79.

Goovaerts, P. (2017b). Monitoring the aftermath of Flint drinking water contamination crisis: Another case of sampling bias? Science of the Total Environment, 590, 139-153.

Goovaerts, P., AvRuskin, G., Meliker, J., Slotnick, M., Jacquez, G., & Nriagu, J. (2005). Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan. *Water Resources Research, 41*(7).

Griffith, D. A. (2004). A spatial filtering specification for the autologistic model. *Environment and planning A, 36*(10), 1791-1811.

Ha, S., Hu, H., Mao, L., Roussos-Ross, D., Roth, J., & Xu, X. (2016). Potential selection bias associated with using geocoded birth records for epidemiologic research. *Annals of epidemiology*, *26*(3), 204-211.

Hanna-Attisha, M., LaChance, J., Sadler, R. C., & Champney Schnepp, A. (2016). Elevated blood lead levels in children associated with the Flint drinking water crisis: a spatial analysis of risk and public health response. *American journal of public health*, *106*(2), 283-290.

Harden, S. L., Chapman, M. J., & Harned, D. A. (2009). *Characterization of Groundwater Quality Based on Regional Geologic Setting in the Piedmont and Blue Ridge Physiographic Provinces, North Carolina*: US Geological Survey.

Hart, T. C., & Zandbergen, P. A. (2013). Reference data and geocoding quality. *Policing: An International Journal of Police Strategies & Management*.

Hassan, M. M., & Atkins, P. J. (2011). Application of geostatistics with Indicator Kriging for analyzing spatial variability of groundwater arsenic concentrations in Southwest Bangladesh. *Journal of Environmental Science and Health, Part A, 46*(11), 1185-1196. doi:10.1080/10934529.2011.598771

He, X., Li, P., Wu, J., Wei, M., Ren, X., & Wang, D. (2020). Poor groundwater quality and high potential health risks in the Datong Basin, northern China: research from published data. *Environmental Geochemistry and Health*, 1-22.

Heck, J. E., Andrew, A. S., Onega, T., Rigas, J. R., Jackson, B. P., Karagas, M. R., & Duell, E. J. (2009). Lung cancer in a US population with low to moderate arsenic exposure. *Environmental health perspectives, 117*(11), 1718-1723.

Heinzerling, A., Stuckey, P. M. J., Scheuer, T., Xu, K., Perkins, K. M., Resseger, H., . . . Acosta, M. (2020). Transmission of COVID-19 to health care personnel during exposures to a hospitalized patient—Solano County, California, February 2020. *MMWR. Morbidity and mortality weekly report*, 69.

Hubbard, R., Newton, G., & Hill, G. (2004). Water quality and the grazing animal. *Journal of animal science, 82*(suppl_13), E255-E263.

Hynds, P. D., Misstear, B. D., & Gill, L. W. (2012). Development of a microbial contamination susceptibility model for private domestic groundwater sources. *Water Resources Research, 48*(12).

International Agency for Research on Cancer. (2004). Some drinking-water disinfectants and contaminants, including arsenic (Vol. 84): IARC.

Jacquez, G. M. (2012). A research agenda: does geocoding positional error matter in health GIS studies? *Spatial and spatio-temporal epidemiology, 3*(1), 7-16.

Jacquez, G. M., & Rommel, R. (2009). Local indicators of geocoding accuracy (LIGA): theory and application. *International Journal of Health Geographics, 8*(1), 60.

James, K. A., Byers, T., Hokanson, J. E., Meliker, J. R., Zerbe, G. O., & Marshall, J. A. (2015). Association between lifetime exposure to inorganic arsenic in drinking water and coronary heart disease in Colorado residents. *Environmental health perspectives, 123*(2), 128-134.

Kaplan, O. B. (2014). *Septic systems handbook*: CRC Press.

Karagas, M. R., Gossai, A., Pierce, B., & Ahsan, H. (2015). Drinking water arsenic contamination, skin lesions, and malignancies: a systematic review of the global evidence. *Current environmental health reports, 2*(1), 52-68.

Kim, D., Miranda, M. L., Tootoo, J., Bradley, P., & Gelfand, A. E. (2011). Spatial modeling for groundwater arsenic levels in North Carolina. *Environmental Science & Technology, 45*(11), 4824-4831. doi:10.1021/es103336s

Knierim, K. J., Hays, P. D., & Bowman, D. (2015). Quantifying the variability in Escherichia coli (E. coli) throughout storm events at a karst spring in northwestern Arkansas, United States. *Environmental earth sciences, 74*(6), 4607-4623.

Knobeloch, L., Gorski, P., Christenson, M., & Anderson, H. (2013). Private drinking water quality in rural Wisconsin. *Journal of Environmental Health, 75*(7), 16-21.

Lee, J. J., Liu, C. W., Jang, C. S., & Liang, C. P. (2008). Zonal management of multi-purpose use of water from arsenic-affected aquifers by using a multi-variable indicator kriging approach. *Journal of hydrology, 359*(3), 260-273. doi:10.1016/j.jhydrol.2008.07.015

Li, X., Li, B., Xi, S., Zheng, Q., Lv, X., & Sun, G. (2013). Prolonged environmental exposure of arsenic through drinking water on the risk of hypertension and type 2 diabetes. *Environmental Science and Pollution Research, 20*(11), 8151-8161.

Lu, Y., & Delmelle, E. (2019). *Geospatial Technologies for Urban Health*: Springer.

Maantay, J. A., & McLafferty, S. (2011). *Geospatial analysis of environmental health* (Vol. 4): Springer Science & Business Media.

MacDonald Gibson, J., & Pieper, K. J. (2017). Strategies to improve private-well water quality: a North Carolina perspective. *Environmental health perspectives, 125*(7), 076001. doi:10.1289/EHP890.

Macintyre, S., Ellaway, A., & Cummins, S. (2002). Place effects on health: how can we conceptualise, operationalise and measure them? *Social science & medicine, 55*(1), 125-139.

Maran, N., Crispim, B., Iahnn, S., Araújo, R., Grisolia, A., & Oliveira, K. (2016). Depth and well type related to groundwater microbiological contamination. *13*(10), 1036.

Maupin, M. A., Kenny, J. F., Hutson, S. S., Lovelace, J. K., Barber, N. L., & Linsey, K. S. (2014). *Estimated use of water in the United States in 2010* (2330-5703). Retrieved from

Mazumdar, S., Rushton, G., Smith, B. J., Zimmerman, D. L., & Donham, K. J. (2008). Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics, 7*(1), 13.

McDonald, Y. J., Schwind, M., Goldberg, D. W., Lampley, A., & Wheeler, C. M. (2017). An analysis of the process and results of manual geocode correction. *Geospatial health, 12*(1), 526.

McQuillan, D. (2004). *Ground-water quality impacts from on-site septic systems.* Paper presented at the Proceedings, National Onsite Wastewater Recycling Association, 13th Annual Conference, Albuquerque, NM.

Meliker, J. R., AvRuskin, G. A., Slotnick, M. J., Goovaerts, P., Schottenfeld, D., Jacquez, G. M., & Nriagu, J. O. (2008). Validity of spatial models of arsenic concentrations in private well water. *Environmental Research, 106*(1), 42-50.

Meliker, J. R., Slotnick, M. J., AvRuskin, G. A., Schottenfeld, D., Jacquez, G. M., Wilson, M. L., . . . Nriagu, J. O. (2010). Lifetime exposure to arsenic in drinking water and bladder cancer: a population-based case–control study in Michigan, USA. *Cancer causes & control, 21*(5), 745-757.

Melo, F., & Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS, 21*(1), 3-38.

Mesner, N. (2012). How to Protect Your Well Water-Homeowner Fact Sheet.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers, 94*(2), 284-289.

Morgan, D. R., Chidi, V., & Owen, R. L. (2015). 49. Gastroenteritis. *Clinical infectious disease*, 334.

Murray, A. T., Grubesic, T. H., Wei, R., & Mack, E. A. (2011). A hybrid geocoding methodology for spatio-temporal data. *Transactions in GIS, 15*(6), 795-809.

National Research Council. (1991). *Environmental Epidemiology, Volume 1: Public Health and Hazardous Wastes* (Vol. 1): National Academies Press.

Navoni, J., De Pietri, D., Olmos, V., Gimenez, C., Mitre, G. B., De Titto, E., & Lepori, E. V. (2014). Human health risk assessment with spatial analysis: study of a population chronically exposed to arsenic through drinking water from Argentina. *Science of the Total Environment, 499*, 166-174.

North Carolina Department of Health and Human Services. (2019, October 1, 2019). Well Water and Health. Retrieved from https://epi.dph.ncdhhs.gov/oee/wellwater/figures.html

Nuckols, J. R., Ward, M. H., & Jarup, L. (2004). Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental health perspectives, 112*(9), 1007-1015.

Nuckols, J. R., Ward, M. H., & Jarup, L. (2004). Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental health perspectives, 112*(9), 1007-1015.

Olabisi, O. E., Awonusi, A. J., & Adebayo, O. J. (2008). Assessment of bacteria pollution of shallow well water in Abeokuta, Southwestern Nigeria. *Life Science Journal, 5*(1), 59-65.

Owusu, C., Baker, K. M., Paul, R., & Curtis, A. B. (2018). Modelling individual vulnerability to sexually transmitted infections to optimise intervention strategies: analysis of surveillance data from Kalamazoo County, Michigan, USA. *Sexually transmitted infections, 94*(5), 353-358.

Owusu, C., Lan, Y., Zheng, M., Tang, W., & Delmelle, E. (2017). Geocoding fundamentals and associated challenges. *Geospatial Data Science Techniques and Applications*, 41-62.

Pandey, P. K., Kass, P. H., Soupir, M. L., Biswas, S., & Singh, V. P. (2014). Contamination of water resources by pathogenic bacteria. *Amb Express, 4*(1), 51.

Patterson, R. A. (1999). *Peat treatment of septic tank effluent.* Paper presented at the Proceedings of On-site.

Pieper, K. J., Krometis, L.-A. H., Gallagher, D. L., Benham, B. L., Edwards, M. J. J. o. w., & health. (2015). Incidence of waterborne lead in private drinking water systems in Virginia. *13*(3), 897-908.

Pippin, C. G. (2005). Distribution of total arsenic in groundwater in the North Carolina Piedmont, paper presented at 2005 NGWA Naturally Occurring Contaminants Conference—Arsenic, Radium, Radon, and Uranium,. Retrieved from http://h2o.enr.state.nc.us/gwp/Arsenic_Studies.htm.

Procopio, N. A., Atherholt, T. B., Goodrow, S. M., & Lester, L. A. (2017). The likelihood of coliform bacteria in NJ domestic wells based on precipitation and other factors. *Groundwater*, *55*(5), 722-735.

Randall, S. M., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, *13*(1), 64.

Reid, J. C., Pippin, C. G., Haven, W. T., & Wooten, R. (2005). *Assessing the Source for Arsenic in Groundwater, North Carolina Piedmont*. Raleigh Retrieved from http://terraquestpc.com/downloads/technical/ArsenicStudy.pdf

Reif, J. S., Burch, J. B., Nuckols, J. R., Metzger, L., Ellington, D., & Anger, W. K. (2003). Neurobehavioral effects of exposure to trichloroethylene through a municipal water supply. *Environmental Research, 93*(3), 248-258.

Ren, X., McHale, C. M., Skibola, C. F., Smith, A. H., Smith, M. T., & Zhang, L. J. E. h. p. (2010). An emerging role for epigenetic dysregulation in arsenic toxicity and carcinogenesis. *119*(1), 11-19.

Rogerson, P. A. (2019). *Statistical methods for geography: a student's guide*: SAGE Publications Limited.

Rosu, A., & Chen, D. (2016). An improved approach for geocoding Canadian postal code–based data in health-related studies. *The Canadian Geographer/Le Géographe canadien, 60*(2), 270-281.

Sadeghi, A. M., & Arnold, J. G. (2002). *A SWAT/microbial sub-model for predicting pathogen loadings in surface and groundwater at watershed and basin scales.* Paper presented at the Total Maximum Daily Load (TMDL): Environmental Regulations, Proceedings of 2002 Conference.

Samadder, S. R., & Subbarao, C. (2007). GIS approach of delineation and risk assessment of areas affected by arsenic pollution in drinking water. *Journal of Environmental Engineering, 133*(7), 742-749.

Sanders, A. P., Messier, K. P., Shehee, M., Rudo, K., Serre, M. L., & Fry, R. C. (2012). Arsenic in North Carolina: public health implications. *Environment international, 38*(1), 10-16. doi:10.1016/j.envint.2011.08.005

Sarkar, A., Krishnapillai, M., & Valcour, J. (2012). A study of groundwater quality of private wells in Western Newfoundland communities.

Schmidlin, K., Clough-Gorr, K. M., & Spoerri, A. (2015). Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. *BMC medical research methodology, 15*(1), 46.

Schneider, A. B., & Sarne, D. H. (2005). Long-term risks for thyroid cancer and other neoplasms after exposure to radiation. *Nature clinical practice Endocrinology & metabolism*, 1(2), 82-91

Sonderman, J. S., Mumma, M. T., Cohen, S. S., Cope, E. L., Blot, W. J., & Signorello, L. B. (2012). A multi-stage approach to maximizing geocoding success in a large population-based cohort study through automated and interactive processes. *Geospatial health, 6*(2), 273.

Steinmaus, C., Yuan, Y., Bates, M. N., & Smith, A. H. (2003). Case-control study of bladder cancer and drinking water arsenic in the western United States. *American journal of epidemiology, 158*(12), 1193-1201.

Swartz, C. H., Reddy, S., Benotti, M. J., Yin, H., Barber, L. B., Brownawell, B. J., & Rudel, R. A. (2006). Steroid estrogens, nonylphenol ethoxylate metabolites, and other wastewater contaminants in groundwater affected by a residential septic system on Cape Cod, MA. *Environmental Science & Technology, 40*(16), 4894-4902.

Swistock, B. R., Clemens, S., Sharpe, W. E., & Rummel, S. (2013). Water quality and management of private drinking water wells in Pennsylvania. *Journal of Environmental Health, 75*(6), 60-67.

Thacker, S. B., Stroup, D. F., Parrish, R. G., & Anderson, H. A. (1996). Surveillance in environmental public health: issues, systems, and sources. *American journal of public health, 86*(5), 633-638.

The Gaston County Board of Health. (2011). *Gaston County Well Ordinance of 1989 - General provisions, definitions, registration, and variance*. Gastonia, NC

Tiemann, M. (2014). *Safe drinking water act (SDWA): a summary of the act and its major requirements*.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(sup1), 234-240.

Tobler, W. R. (1979). Cellular geography. In *Philosophy in geography* (pp. 379-386): Springer.

Uhlmann, S., Galanis, E., Takaro, T., Mak, S., Gustafson, L., Embree, G., . . . Isaac-Renton, J. (2009). Where's the pump? Associating sporadic enteric disease with drinking water using a geographic information system, in British Columbia, Canada, 1996–2005. *Journal of water and health, 7*(4), 692-698.

USDA Natural Resources Conservation Service. (2019). *Gaston County Septic Tank Adsorption Fileds*. Washington, DC Retrieved from https://websoilsurvey.sc.egov.usda.gov/WssProduct/kf3sv5cqxtleqjlc1bjsv34q/DL_00000/20200227_13511801629_1_Soil_Report.pdf

USEPA. (1994, June 10, 2019). Method 200.8: Determination of Trace Elements in Waters and Wastes by Inductively Coupled Plasma-Mass Spectrometry. Retrieved from https://www.epa.gov/sites/production/files/2015-06/documents/epa-200.8.pdf

USEPA. (2015). *Groundwater contamination*. EPA Washington, DC. Retrieved from https://www.epa.gov/sites/production/files/2015-08/documents/mgwc-gwc1.pdf

USEPA. (2017, February 24, 2017). Revised Total Coliform Rule And Total Coliform Rule. Retrieved from https://www.epa.gov/dwreginfo/revised-total-coliform-rule-and-total-coliform-rule#rule-summary

USEPA. (2019a). Information about Public Water Systems. Retrieved from https://www.epa.gov/dwreginfo/information-about-public-water-systems

USEPA. (2019b, November 6, 2019). Learn About Private Water Wells. Retrieved from https://www.epa.gov/privatewells/learn-about-private-water-wells

VanDerwerker, T., Zhang, L., Ling, E., Benham, B., & Schreiber, M. (2018). Evaluating geologic sources of arsenic in well water in Virginia (USA). *International journal of environmental research and public health, 15*(4), 787.

Wallender, E. K., Ailes, E. C., Yoder, J. S., Roberts, V. A., & Brunkard, J. M. (2014). Contributing factors to disease outbreaks associated with untreated groundwater. *Groundwater, 52*(6), 886-897.

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., . . . Hartge, P. (2005). Positional accuracy of two methods of geocoding. *Epidemiology*, 542-547.

Weis, B. K., Balshaw, D., Barr, J. R., Brown, D., Ellisman, M., Lioy, P., . . . Sohn, L. (2005). Personalized exposure assessment: promising approaches for human environmental health research. *Environmental health perspectives, 113*(7), 840-848.

Whitsel, E. A., Rose, K. M., Wood, J. L., Henley, A. C., Liao, D., & Heiss, G. (2004). Accuracy and repeatability of commercial geocoding. *American journal of epidemiology, 160*(10), 1023-1029.

WHO. (2016). 10 facts on preventing disease through healthy environments. Retrieved from https://www.who.int/features/factfiles/environmental-disease-burden/en/

World Health Organization (WHO). (2019). Arsenic. Retrieved from https://www.who.int/news-room/fact-sheets/detail/arsenic

Wu, H., & Huffer, F. R. W. (1997). Modelling the distribution of plant species using the autologistic regression model. *Environmental and ecological Statistics, 4*(1), 31-48.

Yates, M. V. (1985). Septic tank density and ground-water contamination. *Groundwater, 23*(5), 586-591.

Yuan, Y., Marshall, G., Ferreccio, C., Steinmaus, C., Liaw, J., Bates, M., & Smith, A. H. (2010). Kidney cancer mortality: fifty-year latency patterns related to arsenic exposure. *J Epidemiology*, 103-108.

Zandbergen, P. A. (2007). Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health, 7*(1), 37.

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems, 32*(3), 214-232.

Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass, 3*(2), 647-680.

Zandbergen, P. A., & Hart, T. C. (2009). Geocoding accuracy considerations in determining residency restrictions for sex offenders. *Criminal Justice Policy Review, 20*(1), 62-90.

Zhang, J., & Goodchild, M. F. (2002). *Uncertainty in geographical information*: CRC press.

Zimmerman, D. L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics, 64*(1), 262-270. doi:10.1111/j.1541-0420.2007.00870.x

Zimmerman, D. L., & Li, J. (2010). The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *J International journal of health geographics, 9*(1), 10.

Zimmerman, D. L., Fang, X., Mazumdar, S., & Rushton, G. (2007). Modeling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics, 6*(1), 1.

Chapter 2

Bentley, J. P., Ford, J. B., Taylor, L. K., Irvine, K. A., & Roberts, C. L. (2012). Investigating linkage rates among probabilistically linked birth and hospitalization records. *BMC medical research methodology, 12*(1), 149.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., & Freudenheim, J. L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology, 14*(4), 408-412.

Cromley, E. K., & McLafferty, S. L. (2011). *GIS and public health*: Guilford Press.

Fabro, A. Y. R., Ávila, J. G. P., Alberich, M. V. E., Sansores, S. A. C., & Camargo-Valero, M. A. (2015). Spatial distribution of nitrate health risk associated with groundwater use as drinking water in Merida, Mexico. *Applied Geography, 65,* 49-57.

Goldberg, D. W. (2011). Improving Geocoding Match Rates with Spatially-Varying Block Metrics. *Transactions in GIS, 15*(6), 829-850.

Goldberg, D. W., Ballard, M., Boyd, J. H., Mullan, N., Garfield, C., Rosman, D., . . . Semmens, J. B. (2013). An evaluation framework for comparing geocoding systems. *International Journal of Health Geographics, 12*(1), 50.

Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2007). From text to geographic coordinates: the current state of geocoding. *19*(1), 33.

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of hydrology, 228*(1-2), 113-129.

Ha, S., Hu, H., Mao, L., Roussos-Ross, D., Roth, J., & Xu, X. (2016). Potential selection bias associated with using geocoded birth records for epidemiologic research. *26*(3), 204-211.

Harden, S. L., Chapman, M. J., & Harned, D. A. (2009). *Characterization of Groundwater Quality Based on Regional Geologic Setting in the Piedmont and Blue Ridge Physiographic Provinces, North Carolina*: US Geological Survey.

Huggins, F. E., Senior, C. L., Chu, P., Ladwig, K., & Huffman, G. P. (2007). Selenium and arsenic speciation in fly ash from full-scale coal-burning utility plants. *Environmental Science & Technology, 41*(9), 3284-3289.

MacDonald Gibson, J., & Pieper, K. J. (2017). Strategies to improve private-well water quality: a North Carolina perspective. *Environmental health perspectives, 125*(7), 076001. doi:10.1289/EHP890.

McDonald, Y. J., Schwind, M., Goldberg, D. W., Lampley, A., & Wheeler, C. M. (2017). An analysis of the process and results of manual geocode correction. *Geospatial health, 12*(1), 526.

Mecklenburg County Health Department. (2019). Well Information System 4.0. Retrieved from https://edmsmapserver.mecklenburgcountync.gov/wis4/index.html

Messier, K. P., Kane, E., Bolich, R., & Serre, M. L. (2014). Nitrate variability in groundwater of North Carolina using monitoring and private well data models. *Environmental Science & Technology, 48*(18), 10804. doi:10.1021/es502725f

Murray, A. T., Grubesic, T. H., Wei, R., & Mack, E. A. (2011). A hybrid geocoding methodology for spatio-temporal data. *Transactions in GIS, 15*(6), 795-809.

North Carolina Department of Health and Human Services. (2019, October 1, 2019). Well Water and Health. Retrieved from https://epi.dph.ncdhhs.gov/oee/wellwater/figures.html

Owusu, C., Lan, Y., Zheng, M., Tang, W., & Delmelle, E. (2017). Geocoding fundamentals and associated challenges. *Geospatial Data Science Techniques and Applications*, 41-62.

Pippin, C. G. (2005). Distribution of total arsenic in groundwater in the North Carolina Piedmont, paper presented at 2005 NGWA Naturally Occurring Contaminants Conference—Arsenic, Radium, Radon, and Uranium,. Retrieved from http://h2o.enr.state.nc.us/gwp/Arsenic_Studies.htm.

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical reservoir modeling*: Oxford university press.

Randall, S. M., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making, 13*(1), 64.

Schaider, L. A., Ackerman, J. M., & Rudel, R. A. (2016). Septic systems as sources of organic wastewater compounds in domestic drinking water wells in a shallow sand and gravel aquifer. *Science of the Total Environment, 547*, 470-481.

Schmidlin, K., Clough-Gorr, K. M., & Spoerri, A. (2015). Privacy Preserving Probabilistic Record Linkage (P3RL): a novel method for linking existing health-related data and maintaining participant confidentiality. *BMC medical research methodology, 15*(1), 46.

Sonderman, J. S., Mumma, M. T., Cohen, S. S., Cope, E. L., Blot, W. J., & Signorello, L. B. (2012). A multi-stage approach to maximizing geocoding success in a large population-based cohort study through automated and interactive processes. *Geospatial health, 6*(2), 273.

United States Environmental Protection Agency. (2019, April 26, 2019). Private Drinking Water Wells. Retrieved from: https://www.epa.gov/privatewells

Ward, M. H., Nuckols, J. R., Giglierano, J., Bonner, M. R., Wolter, C., Airola, M., . . . Hartge, P. (2005). Positional accuracy of two methods of geocoding. *Epidemiology*, 542-547.

Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass, 3*(2), 647-680.

Zhan, F. B., Brender, J. D., Lima, I. D., Suarez, L., & Langlois, P. H. (2006). Match rate and positional accuracy of two geocoding methods for epidemiologic research. *16*(11), 842-849.

Zimmerman, D. L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics, 64*(1), 262-270. doi:10.1111/j.1541-0420.2007.00870.x

Chapter 3

Almberg, K. S., Turyk, M. E., Jones, R. M., Rankin, K., Freels, S., Graber, J. M., & Stayner, L. T. (2017). Arsenic in drinking water and adverse birth outcomes in Ohio. *Environmental Research, 157,* 52-59. doi:10.1016/j.envres.2017.05.010

Antunes, I. M., & Albuquerque, M. T. (2013). Using indicator kriging for the evaluation of arsenic potential contamination in an abandoned mining area (Portugal). *Science of the Total Environment, 442*, 545-552. doi:10.1016/j.scitotenv.2012.10.010

Ayotte, J. D., Medalie, L., Qi, S. L., Backer, L. C., & Nolan, B. T. (2017). Estimating the high-arsenic domestic-well population in the conterminous United States. *Environmental Science & Technology, 51*(21), 12443-12454. doi:10.1021/acs.est.7b02881

Ayotte, J. D., Montgomery, D. L., Flanagan, S. M., & Robinson, K. W. (2003). Arsenic in groundwater in eastern New England: occurrence, controls, and human health implications. *Environmental Science & Technology, 37*(10), 2075-2083. doi:10.1021/es026211g

Ayotte, J. D., Nolan, B. T., Nuckols, J. R., Cantor, K. P., Robinson, G. R., Baris, D., . . . Silverman, D. T. (2006). Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. *Environmental Science & Technology, 40*(11), 3578-3585. doi:10.1021/es051972f

Benbrahim-Tallaa, L., & Waalkes, M. P. (2008). Inorganic arsenic and human prostate cancer. *Environmental health perspectives, 116*(2), 158-164. doi:10.1289/ehp.10423

Bloom, M. S., Neamtiu, I. A., Surdu, S., Pop, C., Anastasiu, D., Appleton, A. A., . . . Gurzau, E. S. (2016). Low level arsenic contaminated water consumption and birth outcomes in Romania—An exploratory study. *Reproductive Toxicology, 59*, 8-16. doi:10.1016/j.reprotox.2015.10.012

Bo, Y. C., Song, C., Wang, J. F., & Li, X. W. (2014). Using an autologistic regression model to identify spatial risk factors and spatial risk patterns of hand, foot and mouth disease (HFMD) in Mainland China. *BMC Public Health, 14*(1), 358. doi:10.1186/1471-2458-14-358

Bräuner, E. V., Nordsborg, R. B., Andersen, Z. J., Tjønneland, A., Loft, S., & Raaschou-Nielsen, O. (2014). Long-term exposure to low-level arsenic in drinking water and diabetes incidence: a prospective study of the diet, cancer and health cohort. *Environmental health perspectives, 122*(10), 1059-1065. doi:10.1289/ehp.1408198

Bretzler, A., Lalanne, F., Nikiema, J., Podgorski, J., Pfenninger, N., Berg, M., & Schirmer, M. (2017). Groundwater arsenic contamination in Burkina Faso, West Africa: Predicting and verifying regions at risk. *Science of the Total Environment, 584*, 958-970. doi:10.1016/j.scitotenv.2017.01.147

Carter, J. M., Driscoll, D. G., Williamson, J. E., & Lindquist, V. A. (2002). Atlas of water resources in the Black Hills area, South Dakota. Retrieved from https://pubs.usgs.gov/ha/ha747/pdf/definition.pdf

Centers for Disease Control and Prevention (CDC). (2019). Safer Well Water through Stronger Public Health Programs. Retrieved from https://blogs.cdc.gov/yourhealthyourenvironment/2019/11/25/safer-well-water-through-stronger-public-health-programs/

Chapman, M. J., Cravotta, C., Szabo, Z., & Lindsey, B. D. (2013). *Naturally Occurring Contaminants in the Piedmont and Blue Ridge Cystalline-rock Aquifers and Piedmont Early Mesozoic Basin Siliciclastic-rock Aquifers, Eastern United States, 1994-2008*: US Department of the Interior, US Geological Survey.

Daniel, C. C., & Dahlen, P. R. (2002). *Preliminary hydrogeologic assessment and study plan for a regional ground-water resource investigation of the Blue Ridge and Piedmont Provinces of North Carolina* (Vol. 2): US Department of the Interior, US Geological Survey.

Dauphiné, D. C., Smith, A. H., Yuan, Y., Balmes, J. R., Bates, M. N., & Steinmaus, C. (2013). Case-control study of arsenic in drinking water and lung cancer in California and Nevada. *International journal of environmental research and public health, 10*(8), 3310-3324. doi:10.3390/ijerph10083310

Dormann, C. F. (2007). Assessing the validity of autologistic regression. *Ecological modelling, 207*(2-4), 234-242.

Dummer, T., Yu, Z., Nauta, L., Murimboh, J., & Parker, L. (2015). Geostatistical modelling of arsenic in drinking water wells and related toenail arsenic concentrations across Nova Scotia, Canada. *Science of the Total Environment, 505*, 1248-1258. doi:10.1016/j.scitotenv.2014.02.055

Ethan, D., & Xiao-Ming, L. (2018). Examining the Geologic Link of Arsenic Contamination in Groundwater in Orange County, North Carolina. *Frontiers in Earth Science, 6*. doi:10.3389/feart.2018.00111

Evans, J. S., & Ram, K. (2020). Spatial Analysis and Modelling Utilities. 1.3-1. Retrieved from https://github.com/jeffreyevans/spatialEco

Focazio, M. J., Tipton, D., Dunkle Shapiro, S., & Geiger, L. H. (2006). The chemical quality of self-supplied domestic well water in the United States. *Groundwater Monitoring & Remediation, 26*(3), 92-104. doi:10.1111/j.1745-6592.2006.00089.x

Fu, R., Thurman, A. L., Chu, T., Steen-Adams, M. M., & Zhu, J. (2013). On estimation and selection of autologistic regression models via penalized pseudolikelihood. *Journal of agricultural, biological, and environmental statistics, 18*(3), 429-449. doi:10.1007/s13253-013-0144-z

Garelick, H., Jones, H., Dybowska, A., & Valsami-Jones, E. (2009). Arsenic pollution sources. In *Reviews of Environmental Contamination Volume 197* (pp. 17-60). New York, NY: Springer.

Gaus, I., Kinniburgh, D., Talbot, J., & Webster, R. (2003). Geostatistical analysis of arsenic concentration in groundwater in Bangladesh using disjunctive kriging. *Environmental Geology, 44*(8), 939-948. doi:10.1007/s00254-003-0837-7

Goldsmith, R., Milton, D. J., & Horton Jr, J. W. (1988). Geologic map of the Charlotte 1 degrees by 2 degrees Quadrangle, North Carolina and South Carolina. *In Miscellaneous Investigations Series, Report: I-1251-E - U. S. Geological Survey.* Retrieved from https://pubs.usgs.gov/imap/1251e/report.pdf

Goovaerts, P. (2009). AUTO-IK: a 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. *Computers & geosciences*, *35*(6), 1255-1270. doi:10.1016/j.cageo.2008.08.014

Goovaerts, P., AvRuskin, G., Meliker, J., Slotnick, M., Jacquez, G., & Nriagu, J. (2005). Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan. *Water Resources Research, 41*(7). doi:10.1029/2004WR003705

Griffith, D. A. (1987). Spatial autocorrelation. *Resource publications in geography*.

Griffith, D. A. (2004). A spatial filtering specification for the autologistic model. *Environment and Planning A : Economy and Space, 36*(10), 1791-1811. doi:doi.org/10.1068/a36247

Gross, E. L., & Low, D. J. (2013). *Arsenic concentrations, related environmental factors, and the predicted probability of elevated arsenic in groundwater in Pennsylvania*: US Department of the Interior, US Geological Survey.

Gurung, J. K., Ishiga, H., & Khadka, M. S. J. E. G. (2005). Geological and geochemical examination of arsenic contamination in groundwater in the Holocene Terai Basin, Nepal. *49*(1), 98-113.

Hamel, L. (2009). Model assessment with ROC curves. In *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1316-1323): IGI Global.

Harden, S. L., Chapman, M. J., & Harned, D. A. (2009). *Characterization of Groundwater Quality Based on Regional Geologic Setting in the Piedmont and Blue Ridge Physiographic Provinces, North Carolina*: US Geological Survey.

Hassan, M. M., & Atkins, P. J. (2011). Application of geostatistics with Indicator Kriging for analyzing spatial variability of groundwater arsenic concentrations in Southwest Bangladesh. *Journal of Environmental Science and Health, Part A, 46*(11), 1185-1196. doi:10.1080/10934529.2011.598771

He, X., Li, P., Ji, Y., Wang, Y., Su, Z., & Elumalai, V. (2020). Groundwater arsenic and fluoride and associated arsenicosis and fluorosis in China: occurrence, distribution and management. *Exposure and Health*. doi:10.1007/s12403-020-00347-8

Heck, J. E., Andrew, A. S., Onega, T., Rigas, J. R., Jackson, B. P., Karagas, M. R., & Duell, E. J. (2009). Lung cancer in a US population with low to moderate arsenic exposure. *Environmental health perspectives, 117*(11), 1718-1723. doi:10.1289/ehp.0900566

Hengl, T. (2009). *A practical guide to geostatistical mapping* (2nd ed.). Luxembourg: Office for Official Publications of the European Communities.

Hossain, F., & Sivakumar, B. (2006). Spatial pattern of arsenic contamination in shallow wells of Bangladesh: regional geology and nonlinear dynamics. *Stochastic Environmental Research and Risk Assessment, 20*(1), 66-76. doi:10.1007/s00477-005-0012-7

Huebner, M. T., Hatcher, R. D., Jr., & Merschat, A. J. (2017). Confirmation of the southwest continuation of the Cat Square terrane, southern Appalachian Inner Piedmont, with implications for middle Paleozoic collisional orogenesis. *American Journal of Science*, *317*(2), 95-176. doi:10.2475/02.2017.01

James, K. A., Byers, T., Hokanson, J. E., Meliker, J. R., Zerbe, G. O., & Marshall, J. A. (2015). Association between lifetime exposure to inorganic arsenic in drinking water and coronary heart disease in Colorado residents. *Environmental health perspectives, 123*(2), 128-134. doi:10.1289/ehp.1307839

Karagas, M. R., Gossai, A., Pierce, B., & Ahsan, H. (2015). Drinking water arsenic contamination, skin lesions, and malignancies: a systematic review of the global evidence. *Current environmental health reports, 2*(1), 52-68. doi:10.1007/s40572-014-0040-x

Kim, D., Miranda, M. L., Tootoo, J., Bradley, P., & Gelfand, A. E. (2011). Spatial modeling for groundwater arsenic levels in North Carolina. *Environmental Science & Technology, 45*(11), 4824-4831. doi:10.1021/es103336s

Lee, J. J., Liu, C. W., Jang, C. S., & Liang, C. P. (2008). Zonal management of multi-purpose use of water from arsenic-affected aquifers by using a multi-variable indicator kriging approach. *Journal of hydrology, 359*(3), 260-273. doi:10.1016/j.jhydrol.2008.07.015

Li, J., & Heap, A. D. (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics, 6*(3-4), 228-241. doi:10.1016/j.ecoinf.2010.12.003

Li, J., & Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software, 53*, 173-189. doi:10.1016/j.envsoft.2013.12.008

Liu, C., & Sharma, A. (2019). Are you going to get a ticket or a warning for speeding? An autologistic regression analysis in Burlington, VT. *Transportation research interdisciplinary perspectives, 1*, 100001. doi:10.1016/j.trip.2019.100001

Liu, J., Piegorsch, W. W., Grant Schissler, A., & Cutter, S. L. (2018). Autologistic models for benchmark risk or vulnerability assessment of urban terrorism outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 181*(3), 803-823. doi:10.1111/rssa.12323

MacDonald Gibson, J., & Pieper, K. J. (2017). Strategies to improve private-well water quality: a North Carolina perspective. *Environmental health perspectives, 125*(7), 076001-076009. doi:10.1289/EHP890.

Mahram, M., Shahsavari, D., Oveisi, S., & Jalilolghadr, S. (2013). Comparison of hypertension and diabetes mellitus prevalence in areas with and without water arsenic contamination. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences, 18*(5), 408-412.

Meliker, J. R., AvRuskin, G. A., Slotnick, M. J., Goovaerts, P., Schottenfeld, D., Jacquez, G. M., & Nriagu, J. O. (2008). Validity of spatial models of arsenic concentrations in private well water. *Environmental Research, 106*(1), 42-50. doi:10.1016/j.envres.2007.09.001

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers, 94*(2), 284-289. doi:10.1111/j.1467-8306.2004.09402005.x

North Carolina Department of Environmental Quality. (2020). NC Geological Survey. Retrieved from https://deq.nc.gov/about/divisions/energy-mineral-land-resources/north-carolina-geological-survey

North Carolina Department of Health and Human Services. (2020). Minimum detection level for arsenic in NC Division of State Laboratory of Public Health In. Raleigh, NC.

Owusu, C., Lan, Y., Zheng, M., Tang, W., & Delmelle, E. (2017). Geocoding fundamentals and associated challenges. In H. A. Karimi & B. Karimi (Eds.), *Geospatial Data Science Techniques and Applications* (pp. 41-62). Boca Raton, FL: CRC Press.

Pebesma, E., Graeler, B., & Pebesma, M. E. (2019). Package 'gstat'.

Pippin, C. G. (2005). Distribution of total arsenic in groundwater in the North Carolina Piedmont, paper presented at 2005 NGWA Naturally Occurring Contaminants Conference—Arsenic, Radium, Radon, and Uranium,. Retrieved from http://h2o.enr.state.nc.us/gwp/Arsenic_Studies.htm.

Reid, J. C., Pippin, C. G., Haven, W. T., & Wooten, R. (2005). *Assessing the Source for Arsenic in Groundwater, North Carolina Piedmont*. Raleigh Retrieved from http://terraquestpc.com/downloads/technical/ArsenicStudy.pdf

Sanders, A. P., Messier, K. P., Shehee, M., Rudo, K., Serre, M. L., & Fry, R. C. (2012). Arsenic in North Carolina: public health implications. *Environment international, 38*(1), 10-16. doi:10.1016/j.envint.2011.08.005

Seeley, M., Goring, S., & Williams, J. W. (2019). Assessing the environmental and dispersal controls on Fagus grandifolia distributions in the Great Lakes region. *Journal of biogeography, 46*(2), 405-419. doi:10.1111/jbi.13491

Smedley, P. L., & Kinniburgh, D. (2002). A review of the source, behaviour and distribution of arsenic in natural waters. *J Applied geochemistry, 17*(5), 517-568. doi:10.1016/S0883-2927(02)00018-5

Steinmaus, C., Yuan, Y., Bates, M. N., & Smith, A. H. (2003). Case-control study of bladder cancer and drinking water arsenic in the western United States. *American journal of epidemiology, 158*(12), 1193-1201. doi:10.1093/aje/kwg281

Sun, G. (2004). Arsenic contamination and arsenicosis in China. *Toxicology and applied pharmacology, 198*(3), 268-271. doi:10.1016/j.taap.2003.10.017

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(1), 234-240. doi:10.2307/143141

Tobler, W. R. (1979). Cellular geography. In S. Gale & G. Olsson (Eds.), *Philosophy in geography. Theory and Decision Library (An International Series in the Philosophy and Methodology of the Social and Behavioral Sciences)* (Vol. 20, pp. 379-386). Dordrecht: Springer.

Tsuyuki, S. (2008). GIS-based modeling of Javan Hawk-Eagle distribution using logistic and autologistic regression models. *Biological Conservation, 141*(3), 756-769.

USEPA. (1994, June 10, 2019). Method 200.8: Determination of Trace Elements in Waters and Wastes by Inductively Coupled Plasma-Mass Spectrometry. Retrieved from https://www.epa.gov/sites/production/files/2015-06/documents/epa-200.8.pdf

VanDerwerker, T., Zhang, L., Ling, E., Benham, B., & Schreiber, M. (2018). Evaluating geologic sources of arsenic in well water in Virginia (USA). *International journal of environmental research and public health, 15*(4), 787. doi:10.3390/ijerph15040787

Waldron, A. J., Bobyarchick, A. R., Diemer, J., Eppes, M. C., & Meentemeyer, R. (2007). Spatial analysis of factors affecting home radon levels around Moss Lake, NC. *Geological Society of America Abstracts with Programs, 39*(2), 28-29. Retrieved from http://search.proquest.com/docview/50871133/abstract/embedded/DVRAUC63S1IM7RP6?source=fedsrch

Werner, C. K., Bender, J. F., Bobyarchick, A. R., Diemer, J. A., Eppes, M. C., & Waldron, A. S. (2009). Investigation of the source of high radon levels in Cleveland County, North Carolina. *Geological Society of America Abstracts with Programs, 41*(1), 18. Retrieved from http://search.proquest.com/docview/50095280/abstract/embedded/DVRAUC63S1IM7RP6?source=fedsrch

Wu, H., & Huffer, F. R. W. (1997). Modelling the distribution of plant species using the autologistic regression model. *Environmental and ecological Statistics, 4*(1), 31-48. doi:10.1023/A:1018553807765

Yuan, Y., Marshall, G., Ferreccio, C., Steinmaus, C., Liaw, J., Bates, M., & Smith, A. H. (2010). Kidney cancer mortality: fifty-year latency patterns related to arsenic exposure. *J Epidemiology*, 103-108. doi:10.1097/EDE.ObO.13e3181.c21.e46

Chapter 4

Allen, M., Clark, R., Cotruvo, J. A., & Grigg, N. (2018). Drinking Water and Public Health in an Era of Aging Distribution Infrastructure. *23*(4), 301-309.

Anselin, L. (2019). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial Analytical* (pp. 111-126): Routledge.

Beal, C., Gardner, E., & Menzies, N. (2005). Process, performance, and pollution potential: A review of septic tank–soil absorption systems. *Soil Research, 43*(7), 781-802.

Beer, K. D., Gargano, J. W., Roberts, V. A., Hill, V. R., Garrison, L. E., Kutty, P. K., . . . Yoder, J. S. (2015). Surveillance for waterborne disease outbreaks associated with drinking water—United States, 2011–2012. *MMWR. Morbidity and mortality weekly report, 64*(31), 842.

Benedict, K. M., Reses, H., Vigar, M., Roth, D. M., Roberts, V. A., Mattioli, M., . . . Fullerton, K. E. (2017). Surveillance for waterborne disease outbreaks associated with drinking water—United States, 2013–2014. *MMWR. Morbidity and mortality weekly report, 66*(44), 1216.

Bivand, R. (2009). Spatial dependencies: Weighting schemes, statistics and models. R package version 0.4-34. In.

Cabral, J. P. (2010). Water microbiology. Bacterial pathogens and water. *International journal of environmental research and public health, 7*(10), 3657-3703.

Centers for Disease Control and Prevention (CDC).(2009). Well testing. Retrieved from: https://www.cdc.gov/healthywater/drinking/private/wells/testing.html

CDC. (2016, November 10, 2016). Magnitude and burden of Waterborne Disease in the U.S. Retrieved from: https://www.cdc.gov/healthywater/burden/index.html

Charrois, J. W. (2010). Private drinking water supplies: challenges for public health. *Cmaj, 182*(10), 1061-1064.

Conboy, M., & Goss, M. (2000). Natural protection of groundwater against bacteria of fecal origin. *Journal of contaminant hydrology, 43*(1), 1-24.

Fairbrother, J., & Nadeau, E. (2006). Escherichia coli: on-farm contamination of animals. *Rev Sci Tech, 25*(2), 555-569.

Farrell-Poe, K., Jones-McLean, L., & McLean, S. (2010). Microorganisms in Private Water Wells. Retrieved from https://repository.arizona.edu/handle/10150/156925

Gerba, C. P. (2009). Indicator microorganisms. In *Environmental microbiology* (pp. 485-499): Elsevier.

Getis, A. (2010). Spatial autocorrelation. In *Handbook of applied spatial analysis* (pp. 255-278): Springer.

Godfrey, S., Timo, F., & Smith, M. (2006). Microbiological risk assessment and management of shallow groundwater sources in Lichinga, Mozambique. *Water Environment Journal, 20*(3), 194-202.

Gonzales, T. R. (2008). The effects that well depth and wellhead protection have on bacterial contamination of private water wells in the Estes Park Valley, Colorado. *Journal of Environmental Health, 71*(5), 17-23.

Hamel, L. (2009). Model assessment with ROC curves. In *Encyclopedia of Data Warehousing and Mining, Second Edition* (pp. 1316-1323): IGI Global.

Hossain, F., & Sivakumar, B. (2006). Spatial pattern of arsenic contamination in shallow wells of Bangladesh: regional geology and nonlinear dynamics. *Stochastic Environmental Research and Risk Assessment, 20*(1), 66-76. doi:10.1007/s00477-005-0012-7

Hubbard, R., Newton, G., & Hill, G. (2004). Water quality and the grazing animal. *Journal of animal science, 82*(suppl_13), E255-E263.

Hynds, P. D., Misstear, B. D., & Gill, L. W. (2012). Development of a microbial contamination susceptibility model for private domestic groundwater sources. *Water Resources Research, 48*(12).

Kaplan, O. B. (2014). *Septic systems handbook*: CRC Press.

Karathanasis, A., Mueller, T., Boone, B., & Thompson, Y. (2006). Effect of soil depth and texture on fecal bacteria removal from septic effluents. *Journal of water and health, 4*(3), 395-404.

Knobeloch, L., Gorski, P., Christenson, M., & Anderson, H. (2013). Private drinking water quality in rural Wisconsin. *Journal of Environmental Health, 75*(7), 16-21.

Lesnoff, M., Lancelot, R., Lancelot, M. R., & Suggests, M. (2010). Package ‘aod’. In.

Maran, N., Crispim, B., Iahnn, S., Araújo, R., Grisolia, A., & Oliveira, K. (2016). Depth and well type related to groundwater microbiological contamination. *13*(10), 1036.

McQuillan, D. (2004). *Ground-water quality impacts from on-site septic systems.* Paper presented at the Proceedings, National Onsite Wastewater Recycling Association, 13th Annual Conference, Albuquerque, NM.

Mesner, N. (2012). How to Protect Your Well Water-Homeowner Fact Sheet.

National Environmental Services Center. (2019). Septic Stats. Retrieved from http://www.nesc.wvu.edu/septic_idb/northcarolina.htm

Nwachukwu, M. A., Feng, H., Amadi, M. I., & Umunna, F. U. J. W. Q., Exposure. (2010). The causes and the control of selective pollution of shallow wells by coliform bacteria, Imo River Basin Nigeria. *J Water Quality, Exposure Health, 2*(2), 75-84.

Olabisi, O. E., Awonusi, A. J., & Adebayo, O. J. (2008). Assessment of bacteria pollution of shallow well water in Abeokuta, Southwestern Nigeria. *Life Science Journal, 5*(1), 59-65.

Pandey, P. K., Kass, P. H., Soupir, M. L., Biswas, S., & Singh, V. P. (2014). Contamination of water resources by pathogenic bacteria. *Amb Express, 4*(1), 51.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research, 96*(1), 3-14.

Ridpath, A., Taylor, E., Greenstreet, C., Martens, M., Wicke, H., & Martin, C. (2016). Description of calls from private well owners to a national well water hotline, 2013. *Science of the Total Environment, 544*, 601-605.

Sadeghi, A. M., & Arnold, J. G. (2002). *A SWAT/microbial sub-model for predicting pathogen loadings in surface and groundwater at watershed and basin scales.* Paper presented at the Total Maximum Daily Load (TMDL): Environmental Regulations, Proceedings of 2002 Conference.

Sarkar, A., Krishnapillai, M., & Valcour, J. (2012). A study of groundwater quality of private wells in Western Newfoundland communities.

Schaider, L. A., Ackerman, J. M., & Rudel, R. A. (2016). Septic systems as sources of organic wastewater compounds in domestic drinking water wells in a shallow sand and gravel aquifer. *Science of the Total Environment, 547*, 470-481.

Swartz, C. H., Reddy, S., Benotti, M. J., Yin, H., Barber, L. B., Brownawell, B. J., & Rudel, R. A. (2006). Steroid estrogens, nonylphenol ethoxylate metabolites, and other wastewater contaminants in groundwater affected by a residential septic system on Cape Cod, MA. *Environmental Science & Technology, 40*(16), 4894-4902.

Swistock, B. R., Clemens, S., Sharpe, W. E., & Rummel, S. (2013). Water quality and management of private drinking water wells in Pennsylvania. *Journal of Environmental Health, 75*(6), 60-67.

The Gaston County Board of Health. (2011). *Gaston County Well Ordinance of 1989 - General provisions, definitions, registration, and variance*. Gastonia, NC

Toze, S. (1999). PCR and the detection of microbial pathogens in water and wastewater. *Water Research, 33*(17), 3545-3556.

United States Department of Agriculture Natural Resources Conservation Service. (1999). *Soil Taxonomy: A basic system of soil classification for making and interpreting soil surveys*(pp. 886). Retrieved from https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_051232.pdf

United States Geological Survey. (2018). Groundwater Wells. Retrieved from https://www.usgs.gov/special-topic/water-science-school/science/groundwater-wells?qt-science_center_objects=0#qt-science_center_objects

USDA Natural Resources Conservation Service. (2019). *Gaston County Septic Tank Adsorption Fileds*. Washington, DC Retrieved from https://websoilsurvey.sc.egov.usda.gov/WssProduct/kf3sv5cqxtleqjlc1bjsv34q/DL_00000/20200227_13511801629_1_Soil_Report.pdf

USEPA. (2015). *Groundwater contamination*. EPA Washington, DC. Retrieved from https://www.epa.gov/sites/production/files/2015-08/documents/mgwc-gwc1.pdf

USEPA. (2017a). Analytical Methods Approved for Compliance Monitoring under the Revised Total Coliform Rule Retrieved from https://www.epa.gov/sites/production/files/2017-02/documents/rtcr_approved_methods.pdf

USEPA. (2017b, February 24, 2017). Revised Total Coliform Rule And Total Coliform Rule. Retrieved from https://www.epa.gov/dwreginfo/revised-total-coliform-rule-and-total-coliform-rule#rule-summary

USEPA. (2018, December 8, 2018). Septic Systems Overview. Retrieved from https://www.epa.gov/septic/septic-systems-overview

USEPA. (2019, November 6, 2019). Learn About Private Water Wells. Retrieved from https://www.epa.gov/privatewells/learn-about-private-water-wells

Wallender, E. K., Ailes, E. C., Yoder, J. S., Roberts, V. A., & Brunkard, J. M. (2014). Contributing factors to disease outbreaks associated with untreated groundwater. *Groundwater, 52*(6), 886-897.

Yates, M. V. (1985). Septic tank density and ground-water contamination. *Groundwater, 23*(5), 586-591.

Chapter 5

National Research Council. (1991). Environmental Epidemiology, Volume 1: Public Health and Hazardous Wastes (Vol. 1): National Academies Press.

APPENDIX 1: R CODES FOR POSITIONAL ACCURACY ESTIMATION USING KRIGING IN CHAPTER 2

~~~
```{R}
###Created by: Claudio Owusu

##Install required packages##
#install.packages("latticeExtra")
#install.packages("lattice")
#install.packages("sp")
#install.packages("splancs")
#install.packages("rgdal")
#install.packages("gstat")
#install.packages("RColorBrewer")
#install.packages("rgeos")
#install.packages("spatstat")
#install.packages("maptools")
#install.packages("GISTools", dependencies = TRUE)
#install.packages("raster")
#install.packages("tmap")
#install.packages ("constrainedKriging") ## back transforms lognormal krigig
#install.packages ("automap")


##load the required spatial libraies
library(RColorBrewer)
library(latticeExtra)
library(splancs)
library(gstat)
library(rgdal)
library(rgeos)
library(spatstat)
library(maptools)
library(GISTools)
library(raster)
library(tmap)
library(constrainedKriging)
library(fishmethods) ## functions to back transform


##set working directory##
setwd("C:/Users/clowu/Documents/UNCC Dissertation/Chapter
1_Geocoding_Manuscript/Data_For_Analysis/Exposure Misclassification")

##load the required datasets for the analysis
~~~

```
reference <-read.csv ("reference.csv", header=TRUE, sep=",")
parceldata <-read.csv ("parceldataF.csv", header=TRUE, sep=",")
addresspointsdata <-read.csv ("addresspointsdataF.csv", header=TRUE, sep=",")
streetdata <-read.csv ("streetdataF.csv", header=TRUE, sep=",")



## creating a spatial objects from the datasets
coordinates(reference) <- ~xloc +yloc
coordinates(addresspointsdata) <- ~xloc +yloc
coordinates(parceldata) <- ~xloc +yloc
coordinates(streetdata) <- ~xloc +yloc

##load the spatial boundary of Gaston
shp <- readOGR(".", "GC_Boundary")
summary(shp)

#Assign a projection from the boundary shapefile (shp) to all the datasets the datasets

proj4string(addresspointsdata) <- CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
                    +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs")
plot(addresspointsdata, col='red', cex=0.7)

proj4string(parceldata) <- CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
                    +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs")
plot(parceldata, col='black', cex=0.7)

proj4string(streetdata) <- CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
                       +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m
+no_defs")
plot(streetdata, col='grey', cex=0.7)



# A fuction to develop a grid from a dataset with xyz locations
#Npts is the approximate number of points to generate
##This function was borrowed from Dr. Aston Shortbridge##
build.convex.grid <- function (x, y, npts) {
  library(splancs) # for gridding and inout functions
  # First make a convex hull border (splancs poly)
  ch <- chull(x, y) # index for pts on convex hull
  ch <- c(ch, ch[1])
  border <- cbind(x[ch], y[ch])  # This works as a splancs poly
  # Now fill it with grid points
  xy.grid <- gridpts(border, npts)
  return(xy.grid)
}


### Create a surface for prediction and visualization from xyz that approximates Gaston
County ###

cm <- coordinates(reference)
grid <- data.frame(build.convex.grid(cm[,1], cm[,2], 20000))
names(grid) <- c('Xloc', 'Yloc')
gridded(grid) <- ~Xloc+Yloc
plot(grid, add=TRUE, pch=1, cex=0.4) # add this to the points plot

##Assign the same projection in the data to the grid
proj4string(grid) <- CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
                    +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs")


#Convert the Gaston boundary into SpatialPolygons object and make it the same
projection as the data
shp <- shp@polygons
shp <- SpatialPolygons(shp, proj4string=CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
                                    +y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m
+no_defs")) #make sure the shapefile has the same CRS from the data, and from the
prediction grid.
plot(shp)

#Clip the prediction grid with the shapefile
clip_grid <- grid[!is.na(over(grid, shp)),]
plot(clip_grid, add=TRUE, pch=1, cex=0.4)

##CREATE NORTH ARROW AND SCALE BAR FOR THE CHARTS###
l2 = list("SpatialPolygonsRescale", layout.north.arrow(), offset = c(392000,162000),
        scale = 2500)
l3 = list("SpatialPolygonsRescale", layout.scale.bar(), offset = c(390000,160000),
        scale = 5000, fill=c("transparent","black"))
l4 = list("sp.text", c(390000,161000), "0")
l5 = list("sp.text", c(394000,161000), "5000 m")


######1.ANALYSING THE DATA Geocoded using addresspoints#######
```

```
### Variogram analysis
addresspoints_error.vg <- variogram(Error2~1, addresspointsdata, width = 100, cutoff =
2000)

##The output and plot of the observed variogram
addresspoints_error.vg
plot(addresspoints_error.vg)


##Check Anisotrophy using variogram maps
vgm.map1 = variogram(Error2~1, addresspointsdata, width = 100, cutoff = 2000, map =
TRUE)
plot(vgm.map1, threshold = 10)
### Using the Automap to generate the fitting parameters
##library (automap)
#autoAP <- autofitVariogram(Error2~1, addresspointsdata)
#summary(autoAP)


##Choose a best Fitting theoretical variogram using the kappa criteria
options(warn = -1) ##don't print warnings
addresspoints_error.vg.mod <- fit.variogram(addresspoints_error.vg, model=vgm("Sph"))
addresspoints_error.vg.mod
attr(addresspoints_error.vg.mod,"SSEr")
resultsAS<- summary(addresspoints_error.vg.mod)

##Write output to a text
sink("addresspointsmodel.txt")
print(resultsAS)
sink()

#Plot the variogram
png("cex-axis.png")
main<-par(cex.axis= 10.0, cex.lab = 5.0)
plot(addresspoints_error.vg, addresspoints_error.vg.mod, xlab ="Distance (m)",
main=main)
dev.off()



##The predictions
#addresspoints.ok <- krige(Error2~1, addresspointsdata, clip_grid,
model=addresspoints_error.vg.mod, nmax = 5)
#spplot(addresspoints.ok["var1.pred"])
#summary(addresspoints.ok)
```

```
##CHECKING for ANISOTRPY
#addresspoints_error.vg.dir=variogram(Error2~1, addresspointsdata, width = 100,
cutoff = 2000,alpha=c(0,45,90,135))
#addresspoints_error.vg.dir
#plot(addresspoints_error.vg.dir)
#addresspoints_error.vg.mod<-fit.variogram(addresspoints_error.vg,model =
vgm(0.033, "Sph", 650, 0.0192, anis=c(45,0.5)))
#plot(addresspoints_error.vg.dir,addresspoints_error.vg.mod, as.table = TRUE)
#addresspoints_error.vg.mod
#attr(addresspoints_error.vg.mod,"SSEr")


### Predicting the addressPointserrors###
mapcolor <-colorRampPalette(brewer.pal(9, "YlOrRd")) (100)
legendArgs <- list(fun = draw.colorkey,
                   args = list(key = args),
                   corner = c(0.05,.75))
addresspoints.ok <- krige(Error2~1, addresspointsdata, clip_grid,
model=addresspoints_error.vg.mod, nmax=5)
spplot(addresspoints.ok["var1.pred"],
sp.layout=list(l2,l3,l4,l5),col.regions=mapcolor,scales=list(draw=FALSE),
       colorkey = list(space = "right", height = 1.0))+
  layer(sp.polygons(shp, lwd = 1.5))

summary(addresspoints.ok)

##Export output to raster using the rgdal
writeGDAL(addresspoints.ok["var1.pred"], "pred.addresspointsErr2.tif")

##Export variance to raster using the rgdal
writeGDAL(addresspoints.ok["var1.var"], "variance.addresspointsVar2.tif")


######2.ANALYSING THE DATA Geocoded using Parcel#######

### Variogram analysis
parcel_error.vg <- variogram(Error2~1, parceldata, width = 100, cutoff = 2000)
##The output and plot of the observed variogram
parcel_error.vg
plot(parcel_error.vg)

##Check Anisotrophy using variogram maps
vgm.map2 = variogram(Error2~1, parceldata, width = 100, cutoff = 2000, map = TRUE)
plot(vgm.map2, threshold = 10)
```

```
##Choose a best Fitting theoretical variogram using the kappa criteria
options(warn = -1) ##don't print warnings
parcel_error.vg.mod <- fit.variogram(parcel_error.vg, vgm("Sph"))
parcel_error.vg.mod
attr(parcel_error.vg.mod,"SSEr")
resultsAS<- summary(addresspoints_error.vg.mod)

##Write output to a text
sink("parcelsmodel.txt")
print(resultsAS)
sink()

plot(parcel_error.vg, parcel_error.vg.mod, xlab ="Distance (m)")##, cex.axis= 2.0,
cex.lab = 2.0)


### Predicting the parcelserrors###
mapcolor <-colorRampPalette(brewer.pal(9, "YlOrRd")) (100)
legendArgs <- list(fun = draw.colorkey,
                   args = list(key = args),
                   corner = c(0.05,.75))
parcel.ok <- krige(Error2~1, parceldata, clip_grid, model=parcel_error.vg.mod, nmax=5)
spplot(parcel.ok["var1.pred"],
sp.layout=list(l2,l3,l4,l5),col.regions=mapcolor,scales=list(draw=FALSE),
       colorkey = list(space = "right", height = 1.0))+
  layer(sp.polygons(shp, lwd = 1.5))

summary(parcel.ok)

##Export output to raster using the rgdal
writeGDAL(parcel.ok["var1.pred"], "pred.parcelErr2.tif")

##Export variance to raster using the rgdal
writeGDAL(parcel.ok["var1.var"], "variance.parcelVar2.tif")


######3.ANALYSING THE DATA Geocoded using street#######

### Variogram analysis
street_error.vg <- variogram(Error2~1, streetdata, width = 100, cutoff = 2000)

##The output and plot of the observed variogram
street_error.vg
plot(street_error.vg)
```

```
##Check Anisotrophy using variogram maps
vgm.map = variogram(Error2~1, streetdata, width = 100, cutoff = 2000, map = TRUE)
plot(vgm.map, threshold = 10)


##Choose a best Fitting theoretical variogram using the kappa criteria
options(warn = -1) ##don't print warnings
street_error.vg.mod <-fit.variogram(street_error.vg,vgm("Sph"))
street_error.vg.mod
attr(street_error.vg.mod,"SSEr")
resultsAS<- summary(street_error.vg.mod)
##Write output to a text
sink("streetmodel.txt")
print(resultsAS)
sink()


jpeg(file="streetMod.jpg",bg="white", res=300, pointsize = 16, width = 1200, height =
1200, quality = 100)

plot(street_error.vg, plot.number=F, model = street_error.vg.mod, ylim=c(0.04, 0.08), col
="black", cex.axis = 1.5)
##Plot semivariogram
plot(street_error.vg, street_error.vg.mod, xlab ="Distance (m)", cex.axis= 0.7, cex.lab =
1.5, font.axis = 3)


### Predicting the streeterrors###
### Predicting the parcelserrors###
mapcolor <-colorRampPalette(brewer.pal(9, "YlOrRd")) (100)
legendArgs <- list(fun = draw.colorkey,
                   args = list(key = args),
                   corner = c(0.05,.75))
parcel.ok <- krige(Error2~1, parceldata, clip_grid, model=parcel_error.vg.mod, nmax=5)
spplot(parcel.ok["var1.pred"],
sp.layout=list(l2,l3,l4,l5),col.regions=mapcolor,scales=list(draw=FALSE),
       colorkey = list(space = "right", height = 1.0))+
  layer(sp.polygons(shp, lwd = 1.5))

summary(parcel.ok)


### Predicting the streeterrors###
mapcolor <-colorRampPalette(brewer.pal(9, "YlOrRd")) (100)
legendArgs <- list(fun = draw.colorkey,
                   args = list(key = args),
```

~~~
             corner = c(0.05,.75))
street.ok <- krige(Error2~1, streetdata, clip_grid, model=street_error.vg.mod, nmax=5)
spplot(street.ok["var1.pred"],
sp.layout=list(l2,l3,l4,l5),col.regions=mapcolor,scales=list(draw=FALSE),
     colorkey = list(space = "right", height = 1.0))+
  layer(sp.polygons(shp, lwd = 1.5))


##Export output to raster using the rgdal
streetPre<-writeGDAL(street.ok["var1.pred"], "pred.streetErr2.tif")
#StreetPre2 <-raster(streetPre)
#streetpre3 <-bt.log()

##Export variance to raster using the rgdal
writeGDAL(street.ok["var1.var"], "variance.streetVar2.tif")
```
~~~

APPENDIX 2: R CODES FOR AUTOLOGISTIC MODEL FOR CHAPTER 3

~~~
##ARSENIC DATA ANALYSIS
```{r}
#install.packages("ggplot2")
#install.packages("sp")
#install.packages("lctools")
#install.packages("spdep")
#install.packages("spatialEco")
library(spatialEco)
library(sp)
library(ggplot2)
library(spdep)
library(ggpubr)
library(caret)
theme_set(theme_pubr())

```



```{r}
setwd("C:/Users/clowu/Documents/UNCC Dissertation/Chapter
2_Arsenic/Arsenic_Analysis/StatisticalModels_New")
mydata <- read.csv("Final_Arsenic_DataNew2.csv", sep = ',')
colnames(mydata)

##select only the variables important for the modelling
ArsenicData = subset(mydata, select = c(1,14,27:37))
head(ArsenicData)

###transform the data into factors
ArsenicData$Arsenic_Detect2 = as.factor(ArsenicData$Arsenic_Detect2)
levels(ArsenicData$Arsenic_Detect2) = c('No','Yes')
colnames(ArsenicData)

###transform categorical variables into factors
ArsenicData$Depth = as.factor(ArsenicData$Depth)
ArsenicData$Bedrock = as.factor(ArsenicData$Bedrock)
ArsenicData$BedrockNew = as.factor(ArsenicData$BedrockNew)
attach(ArsenicData)
head(ArsenicData)
colnames(ArsenicData)

```

##Summary table
~~~

```{r}
summaryvar = summary(ArsenicData)
summaryvar
write.csv(summaryvar, file = "summary.csv")
```

Plotting histograms of the variables

```{r}

pHplot = ggplot(ArsenicData, aes(x=pH))+ geom_histogram(binwidth=1, bins = 14, color="darkblue", fill="skyblue3")+ scale_x_continuous(name="pH Level", breaks = c(0,2, 4, 6, 8, 10, 12,14)) + scale_y_continuous(name="Frequency")+ theme_pubclean()

pHplot


Rocktypeplot = ggplot(ArsenicData, aes(Bedrock)) + geom_bar(color="darkblue", fill="skyblue3") + theme_pubclean() + labs(y="Frequency")

Rocktypeplot


Depthplot = ggplot(ArsenicData, aes(Depth)) + geom_bar(color="darkblue", fill="skyblue3") + theme_pubclean() + labs(y="Frequency")

Depthplot


colnames(ArsenicData)
```

Model Development

```{r}
##Ordinary logistic regression
lmodel =logistic.regression(ArsenicData, y = 'Arsenic_Detect', x = c('BedrockNew','Depth', 'pH'), penalty = TRUE)
lmodel$model
lmodel$diagTable
lmodel$coefTable
lmodel_pred = predict(lmodel$model, type = 'fitted.ind')


##Spatial autologistic regression
coordinates(ArsenicData) = ~xloc + yloc
```

~~~
lmodel2 = logistic.regression(ArsenicData, y = 'Arsenic_Detect', x =
c('BedrockNew','pH','Depth'), autologistic = TRUE, coords =
coordinates(ArsenicData),longlat = FALSE, penalty = TRUE)

lmodel2$model
   lmodel2$diagTable
     lmodel2$coefTable
     lmodel2$bandwidth

lmodel2_pred = predict(lmodel2$model, type = 'fitted.ind')
autocovariate = lmodel2$AutoCov
residuals2 = lmodel2$Residuals


##write results to csv for mapping
model_arsenic = data.frame(DataID,lmodel2_pred, Arsenic_Detect,autocovariate,
residuals2, xloc, yloc)

write.csv(model_arsenic, file = "model_arsenic.csv")

```


##ROC Curve
```{r}
#install.packages("ROCR")
library(ROCR)
library(pROC)

arsenic_chk = ArsenicData$Arsenic_Detect2

#par(pty ="s")

#roc(arsenic_chk, lmodel_pred, plot = TRUE, legacy.axes = TRUE, xlab = "1-specificity
(False positive rate)", ylab = "Sensitivity (True positive rate)", col="#de2d26", lwd=1,
print.auc = TRUE)

#plot.roc(arsenic_chk, lmodel2_pred,col="#377eb8", lwd=1, print.auc = TRUE, add=
TRUE, print.auc.y=0.4)
#legend("bottomright", legend = c("non-spatial", "spatial"), col =
c("#de2d26","#377eb8" ), lwd = 1)

##Area Under the Receiver Operator Characteristic Curve (AUROC)
chkroc1 = pROC::roc(ArsenicData$Arsenic_Detect2,lmodel2_pred)
chkroc1
ci.auc(chkroc1)
~~~

~~~
##Checking accuracy of predictions
lmodel2_pred0 = rep("No",990)
lmodel2_pred0[lmodel2_pred>.5] ="Yes"
table(lmodel2_pred0, ArsenicData$Arsenic_Detect2)
mean(lmodel2_pred0==ArsenicData$Arsenic_Detect2)

summary(ArsenicData$Arsenic_Detect2)
##Checking the mean squared errors for the testing data set
n0=length(ArsenicData$Arsenic_Detect)
sse10 = sum((ArsenicData$Arsenic_Detect - lmodel2_pred)^2)
mse10 = sse10 / (n0 - 2)
mse10


```
~~~

APPENDIX 3: R CODES FOR KRIGING MAPS IN CHAPTER 3

~~~
###KRIGING ANALYSIS###
```{r}
##Install required packages##
#install.packages("latticeExtra")
#install.packages("lattice")
#install.packages("splancs")
#install.packages("rgdal")
#install.packages("gstat")
#install.packages("RColorBrewer")
#install.packages("rgeos")
#install.packages("spatstat")
#install.packages("maptools")
#install.packages("GISTools", dependencies = TRUE)
#install.packages("raster")
#install.packages("tmap")
#install.packages("sf")

##load the required spatial libraies
library(RColorBrewer)
library(latticeExtra)
library(splancs)
library(gstat)
library(rgdal)
library(rgeos)
library(spatstat)
library(maptools)
library(GISTools)
library(raster)
library(tmap)
library(sf)

```




**3.Spatial autologistic regression probs interpolation **
```{r}

modelresults = read.csv("model_arsenic.csv", sep = ",")

## creating a spatial objects from the datasets
coordinates(modelresults) <- ~xloc +yloc
~~~

```
setwd("C:/Users/clowu/Documents/UNCC Dissertation/Chapter
2_Arsenic/Arsenic_Analysis/StatisticalModels_New/Kriging")
##load the required datasets for the analysis
addresspoints <-read.csv ("addresspoints.csv", header=TRUE, sep=",")

## creating a spatial objects from the datasets
coordinates(addresspoints) <- ~xloc +yloc


#Assign a projection from the boundary shapefile (shp) to all the datasets the datasets
proj4string(modelresults) <- CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
+y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs")

##load the spatial boundary of Gaston
shp <- readOGR(".", "GC_BoundaryF")
summary(shp)

# A fuction to develop a grid from a dataset with xyz locations
#Npts is the approximate number of points to generate
build.convex.grid <- function (x, y, npts) {
  library(splancs) # for gridding and inout functions
  # First make a convex hull border (splancs poly)
  ch <- chull(x, y) # index for pts on convex hull
  ch <- c(ch, ch[1])
  border <- cbind(x[ch], y[ch])  # This works as a splancs poly

  # Now fill it with grid points
  xy.grid <- gridpts(border, npts)
  return(xy.grid)
}


### Create a surface for prediction and visualization from xyz that approximates Gaston
County ###
cm <- coordinates(addresspoints)
grid <- data.frame(build.convex.grid(cm[,1], cm[,2], 10000))
names(grid) <- c('Xloc', 'Yloc')
gridded(grid) <- ~Xloc+Yloc
plot(modelresults, col='blue', cex=0.7)
plot(grid, add=TRUE, pch=1, cex=0.4) # add this to the points plot



##Assign the same projection in the data to the grid
```

~~~
proj4string(grid) <- CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
+y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs")


#Convert the Gaston boundary into SpatialPolygons object and make it the same
projection as the data
shp <- shp@polygons
shp <- SpatialPolygons(shp, proj4string=CRS("+proj=lcc +lat_1=34.33333333333334
+lat_2=36.16666666666666 +lat_0=33.75 +lon_0=-79 +x_0=609601.2192024384
+y_0=0 +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +units=m +no_defs")) #make sure the
shapefile has the same CRS from the data, and from the prediction grid.
plot(shp)

#Clip the prediction grid with the shapefile
clip_grid <- grid[!is.na(over(grid, shp)),]
plot(clip_grid, add=TRUE, pch=1, cex=0.4)

##CREATE NORTH ARROW AND SCALE BAR FOR THE CHARTS###
l2 = list("SpatialPolygonsRescale", layout.north.arrow(), offset = c(392000,162000),
        scale = 2500)
l3 = list("SpatialPolygonsRescale", layout.scale.bar(), offset = c(390000,160000),
        scale = 5000, fill=c("transparent","black"))
l4 = list("sp.text", c(390000,161000), "0")
l5 = list("sp.text", c(394000,161000), "5000 m")

```



#Semivariogram for predicted probabilities in the model
```{r}
modelresults$probs <- (modelresults$lmodel2_pred) * 1
PredMod<- variogram(probs~1, modelresults,boundaries = seq(0,20000, l=51))

##The output and plot of the observed variogram
PredMod
plot(PredMod)


PredMod.vg.mod= fit.variogram(PredMod, vgm(c("Gau", "Exp","Sph")))
PredMod.vg.mod

plot(PredMod, model=PredMod.vg.mod)

```
~~~

~~~
```{r}
### Local neighborhood Indicator OK on high counts of positive test of coliform ###
mapcolor <-colorRampPalette(brewer.pal(9, "YlOrRd")) (100)
legendArgs <- list(fun = draw.colorkey,
                   args = list(key = args),
                   corner = c(0.05,.75))
SpatialAuto.pred <- krige(probs~1, modelresults, clip_grid, model=PredMod.vg.mod,
nmax = 3)
spplot(SpatialAuto.pred["var1.pred"],
sp.layout=list(l2,l3,l4,l5),col.regions=mapcolor,scales=list(draw=FALSE),
       colorkey = list(space = "right", height = 1.0), at = seq(0, 1, .01))+
  layer(sp.polygons(shp, lwd = 1.5))

summary(SpatialAuto.pred)

##Export output to raster using the rgdal
writeGDAL(SpatialAuto.pred["var1.pred"], "pred.SpatialLogNew.tif")

##Export variance to raster using the rgdal
writeGDAL(SpatialAuto.pred["var1.var"], "variance.SpatialLogNew.tif")

```



#Semivariogram for autocovariate variable in the model
```{r}
modelresults$autocov <- (modelresults$autocovariate) * 1
autocovMod<- variogram(autocov~1, modelresults,boundaries = seq(0,20000, l=51))

##The output and plot of the observed variogram
autocovMod
plot(autocovMod)


autocovMod.vg.mod= fit.variogram(autocovMod, vgm(c("Gau", "Exp","Sph")))
autocovMod.vg.mod

plot(autocovMod, model=autocovMod.vg.mod)

```


```{r}
### Predicting the probabilities ###
mapcolor <-colorRampPalette(brewer.pal(9, "YlOrRd")) (100)
legendArgs <- list(fun = draw.colorkey,
                   args = list(key = args),
~~~

~~~
                corner = c(0.05,.75))
autocov.pred <- krige(autocov~1, modelresults, clip_grid, model=autocovMod.vg.mod,
nmax = 3)
spplot(autocov.pred["var1.pred"],
sp.layout=list(l2,l3,l4,l5),col.regions=mapcolor,scales=list(draw=FALSE),
      colorkey = list(space = "right", height = 1.0), at = seq(0, 1, .01))+
  layer(sp.polygons(shp, lwd = 1.5))

summary(autocov.pred)

##Export output to raster using the rgdal
writeGDAL(autocov.pred["var1.pred"], "pred.autocovNew.tif")

##Export variance to raster using the rgdal
writeGDAL(autocov.pred["var1.var"], "variance.autocovNew.tif")

```



#Semivariogram for residuals in the model
```{r}
modelresults$residualterm <- (modelresults$res)
residualtermMod<- variogram(residualterm~1, modelresults,boundaries = seq(0,20000,
l=51))

##The output and plot of the observed variogram
residualtermMod
plot(residualtermMod)


residualtermMod.vg.mod= fit.variogram(residualtermMod, vgm(c("Gau", "Exp","Sph")))
residualtermMod.vg.mod

plot(residualtermMod, model=residualtermMod.vg.mod)

```


##Ordinary kriging of the model residuals
```{r}

mapcolor <-colorRampPalette(brewer.pal(9, "YlOrRd")) (100)
legendArgs <- list(fun = draw.colorkey,
                args = list(key = args),
                corner = c(0.05,.75))
residualterm.pred <- krige(residualterm~1, modelresults, clip_grid,
model=residualtermMod.vg.mod, nmax = 5)
~~~

~~~
spplot(residualterm.pred["var1.pred"],
sp.layout=list(l2,l3,l4,l5),col.regions=mapcolor,scales=list(draw=FALSE),
       colorkey = list(space = "right", height = 1.0), at = seq(0, 1, .01))+
  layer(sp.polygons(shp, lwd = 1.5))

summary(residualterm.pred)

##Export output to raster using the rgdal
writeGDAL(residualterm.pred["var1.pred"], "pred.residualterm.tif")

##Export variance to raster using the rgdal
writeGDAL(residualterm.pred["var1.var"], "variance.residualterm.tif")

```
~~~

APPENDIX 4: R CODES FOR MULTIVARIATE LOGISTIC REGRESSION MODELS IN CHAPTER 4

~~~
#LOGISTIC REGRESSION
##COLIFORM BACTERIA DATA ANALYSIS

```{r}
#install.packages("sjPlot")
#install.packages("tidyr")
#install.packages("caret")
#install.packages("jtools")
#install.packages("ggsci")
#install.packages("pROC")
#install.packages("ggpubr")
library(ggpubr)
library(pROC)
library(ggsci)
library(jtools)
library(caret)
library(tidyr)
library(sjPlot)
library(sjmisc)
library(sjlabelled)

#install.packages("car")
#install.packages("aod")
#install.packages("ggplot2")
#install.packages("sp")
#install.packages("lctools")
#install.packages("spdep")
#install.packages("spatialEco")
#install.packages("caret")
library(caret)
library(aod)
library(ggplot2)
library(sp)
#library(lctools)
library(spdep)
library(spatialEco)
library(car)
```


```{r}
##Set the working directory & read the file
setwd("C:/Users/clowu/Documents/UNCC Dissertation/Chapter 3_Total
Coliform/Coliform_Analysis/StatisticalModelNew")
~~~

~~~
#setwd("E:/StatisticalModelNew")
mydataAll <- read.csv("PathogenData.csv", sep = ',')
colnames(mydataAll)

##select only the variables important for the modelling
PathogenData = subset(mydataAll, select = c(1,19:30))
attach(PathogenData)
head(PathogenData)

###transform categorical data into factors
##1. pathogen
PathogenData$Pathogen2 <-as.factor(PathogenData$Pathogen2)
levels(PathogenData$Pathogen2 )<-c('No','Yes')
head(PathogenData)

##2.welltype
PathogenData$WellType = as.factor(PathogenData$WellType)
head(PathogenData)

##3. Septic Tank Absorption Field
PathogenData$SepTankAF = as.factor(PathogenData$SepTankAF)
head(PathogenData)

##4. MUSYM
PathogenData$MUSYM = as.factor(PathogenData$MUSYM)
head(PathogenData)

PathogenData$CatWellDepth = cut(WellDepth, breaks = c(0, 150, 300, 1025 ), labels = c("1", "2", "3"))

PathogenData$CatWellDepth = as.factor(PathogenData$CatWellDepth)

#cut(WellDepth, breaks = c(0, 150, 300, 1025 ), labels = c("1", "2", "3"))

PathogenData$RatioDepthCasing1 = (CasingDepth/WellDepth) ## ratio of the casing to the well depth

attach(PathogenData)
```


**Summary and correlation tables**
```{r}
summaryvar = summary(PathogenData)
summaryvar
#write.csv(summaryvar, file = "summary.csv")
~~~

~~~
colnames(PathogenData)
correlationtable = round (cor(PathogenData[,c(5:9)]), 3)
correlationtable
#write.csv(correlationtable, file = "correlationtable.csv")

```

**Plotting histograms of the variables**
```{r}
par(mfrow = c(3,2))

hist(Age, xlab = "Age (years)", main = " Histogram of age of well", col='skyblue3')

hist(WellDepth , xlab = "Well Depth(ft)", main = " Histogram of Well depth",
col='skyblue3')

hist(ParcelSize, xlab = "parcel size (acres)", main = " Histogram of parcel size (acres)",
col='skyblue3')

hist(CasingDepth, xlab = "Casing depth", main = " Histogram of Casing depth",
col='skyblue3')

#barplot(prop.table(table(WellType)))
#ggplot(mydata, aes(x=WellType))
```

**perform significant testing categorical variables**
```{r}
##1.Well type
#create a contigency table
WellType_Chisq = table(PathogenData$WellType, PathogenData$Pathogen2)
WellType_Chisq
#Chi-squared test
chisq.test(WellType_Chisq)

##2.Septic tank absorption field rating from USDA
#create a contigency table
SepTankAF_Chisq = table(PathogenData$SepTankAF, PathogenData$Pathogen2)
SepTankAF_Chisq
#Chi-squared test
chisq.test(SepTankAF_Chisq)

#create a contigency table
#PathogenData$CatWellDepth = as.factor(PathogenData$CatWellDepth)
#CatWellType_Chisq = table(PathogenData$CatWellDepth, PathogenData$Pathogen2)
~~~

~~~
#CatWellType_Chisq
#Chi-squared test
#chisq.test(CatWellType_Chisq)
```


**perform siginificant independent t-test of means for continuous variables**
```{r}
##perform Welch Two Sample t-test of continuous variables
t.test(WellDepth ~ Pathogen1, data = PathogenData)
t.test(RatioDepthCasing1 ~ Pathogen1, data = PathogenData)
t.test(Age ~ Pathogen1, data = PathogenData)
t.test(ParcelSize ~ Pathogen1, data = PathogenData)
```

```{r}
##randomized the data sample
#set.seed(123468)
set.seed(123689)
##split the data into partition
partitionRule <- createDataPartition(PathogenData$Pathogen2, p = 0.8, list = F)
trainingSet <- PathogenData[partitionRule,]
testingSet <- PathogenData[-partitionRule,]

summary(trainingSet)
```

**Using the the logit model**
```{r}
##Prediction with logistic regression
model1 = glm(Pathogen1 ~ WellType + RatioDepthCasing1 + Age + ParcelSize +
SepTankAF, data=trainingSet, family = "binomial")
summary(model1)
tab_model(model1)

round(exp(coef(model1)),3) ## odds ratios only
round(exp(confint(model1)),3) ##CI for odds ratio

summary(model1)$coef
model1.probs = predict(model1, type = "response") ##predicted probabilities
AIC(model1)
resi.model1 = residuals(model1)

##Check multicolinearity
viftable =round(vif(model1), 3)
viftable
~~~

~~~
write.csv(viftable, file = "viftable.csv")
```

**Model Validation**
```{r}
##for testing set
summary(testingSet)
model1.probs = as.numeric(unlist(predict(model1, testingSet,type = "response")))

##accuracy check for testing data set
model1.pred = rep("No",231)
model1.pred[model1.probs>.5] ="Yes"
table(model1.pred, testingSet$Pathogen2)
mean(model1.pred==testingSet$Pathogen2)

##Area Under the Receiver Operator Characteristic Curve (AUROC)
chkroc1_test = pROC::roc(testingSet$Pathogen2,model1.probs)
chkroc1_test


###For training set
summary(trainingSet)
model1.probs2 = as.numeric(unlist(predict(model1, trainingSet,type = "response")))

##accuracy check for testing data set
model1.pred2 = rep("No",932)
model1.pred2[model1.probs2<.5] ="Yes"
table(model1.pred2, trainingSet$Pathogen2)
mean(model1.pred2==trainingSet$Pathogen2)

##Area Under the Receiver Operator Characteristic Curve (AUROC)
chkroc1_train = pROC::roc(trainingSet$Pathogen2,model1.probs2)
chkroc1_train

residual.model1 = as.numeric(unlist(residuals(model1)))

```

**Model diagnostics**
```{r}

##Make predictions for total data sets
model1.probsF = as.numeric(unlist(predict(model1, PathogenData,type = "response")))

resi.Model1 = PathogenData$Pathogen1 - model1.probsF
~~~

~~~~
model1_diagnostics = data.frame(SortID,model1.probsF,Pathogen1, Pathogen2,
resi.Model1,WellType, WellDepth, CasingDepth,RatioDepthCasing1, Age, SepTankAF,
ParcelSize, xCoord, yCoord)

write.csv(model1_diagnostics, file = "model1_diagnostics.csv")
```


##plot the prediction with significant variables
```{r}
ModelPred = read.csv("model1_diagnostics.csv",sep = ",", header = TRUE)
attach(ModelPred)

##probablities and well type
welltype_boxplot = ggplot(ModelPred, aes(x=WellType, y= model1.probsF))+
geom_boxplot(aes(fill=WellType)) + scale_color_gradientn(colors = c("#00AFBB",
"#E7B800", "#FC4E07"))+ theme_classic()+theme(legend.position = "none") +
theme(legend.position = "none")+ labs(y="")
welltype_boxplot

Age_pointplot = ggplot(ModelPred, aes(x=Age, y=model1.probsF, colour =
model1.probsF))+ geom_point(size = 3, alpha = 0.6)+ scale_color_gradientn(colors =
c("#00AFBB", "#E7B800", "#FC4E07"))+ theme_classic()+theme(legend.position =
"none") + theme(legend.position = "top")+ labs(y="")
Age_pointplot

RatioDepthCasing1_pointplot = ggplot(ModelPred, aes(x=RatioDepthCasing1,
y=model1.probsF, colour = model1.probsF))+ geom_point(size = 3, alpha = 0.6)+
scale_color_gradientn(colors = c("#00AFBB", "#E7B800", "#FC4E07"))+
theme_classic()+theme(legend.position = "none") + theme(legend.position = "top")+
labs(y="")
RatioDepthCasing1_pointplot


```



##plot the model residuals
```{r}
residmodel1 = read.csv("model1_diagnostics.csv")
summary(residmodel1)
ggplot(residmodel1, aes(xCoord, yCoord, colour =resi.Model1 )) +
  viridis::scale_color_viridis()+
  geom_point(size = 3)
```

***2. FOR DRILLED WELLS***
~~~~

~~~~
```{r}
##Set the working directory & read the file
setwd("C:/Users/clowu/Documents/UNCC Dissertation/Chapter 3_Total Coliform/Coliform_Analysis/StatisticalModelNew/DrilledWells")

#setwd("E:/StatisticalModelNew/DrilledWells")
myDrilledWells <- read.csv("DrilledWells.csv", sep = ',')
colnames(myDrilledWells)

##select only the variables important for the modelling
myDrilledWells = subset(myDrilledWells, select = c(1,18:29))
attach(myDrilledWells)
head(myDrilledWells)

###transform categorical data into factors
##1. pathogen
myDrilledWells$Pathogen2 <-as.factor(myDrilledWells$Pathogen2)
levels(myDrilledWells$Pathogen2 )<-c('No','Yes')
head(myDrilledWells)

##2.welltype
#PathogenData$WellType = as.factor(PathogenData$WellType)
#head(PathogenData)

##3. Septic Tank Absorption Field
myDrilledWells$SepTankAF = as.factor(myDrilledWells$SepTankAF)
head(myDrilledWells)


myDrilledWells$RatioDepthCasing2 = CasingDepth/WellDepth## ratio of the casing to the well depth

attach(myDrilledWells)
```


**Summary and correlation tables**
```{r}
summaryvar = summary(myDrilledWells)
summaryvar
#write.csv(summaryvar, file = "summary.csv")

colnames(myDrilledWells)
correlationtable = round (cor(myDrilledWells[,c(5:9)]), 3)
correlationtable
#write.csv(correlationtable, file = "correlationtable.csv")
```
~~~~

~~~~
**perform significant testing categorical variables**
```{r}
##1.Septic tank absorption field rating from USDA
#create a contigency table
SepTankAF_Chisq = table(myDrilledWells$SepTankAF, myDrilledWells$Pathogen2)
SepTankAF_Chisq
#Chi-squared test
chisq.test(SepTankAF_Chisq)

```


**perform siginificant independent t-test of means for continuous variables**
```{r}
##perform Welch Two Sample t-test of continuous variables
t.test(RatioDepthCasing2 ~ Pathogen1, data = myDrilledWells)
t.test(Age ~ Pathogen1, data = myDrilledWells)
t.test(ParcelSize ~ Pathogen1, data = myDrilledWells)

```



```{r}
##randomized the data sample
#set.seed(123468)
set.seed(123689)
##split the data into partition
partitionRule <- createDataPartition(myDrilledWells$Pathogen2, p = 0.8, list = F)
trainingSet1 <- myDrilledWells[partitionRule,]
testingSet1 <- myDrilledWells[-partitionRule,]

summary(trainingSet1)
```


**Aply Prior and bias correction becasue of small number of events than non-events**
```{r}
model2<- glm(Pathogen1 ~ RatioDepthCasing2 + Age + ParcelSize + SepTankAF,
data=trainingSet1, family = "binomial")
summary(model2)
tab_model(model2)

round(exp(coef(model2)),3) ## odds ratios only
round(exp(confint(model2)),3) ##CI for odds ratio

summary(model1)$coef
~~~~

~~~~
model2.probs = predict(model2, type = "response") ##predicted probabilities
AIC(model2)
resi.model2 = residuals(model2)

##Check multicolinearity
viftable2 =round(vif(model2), 3)
viftable2
write.csv(viftable2, file = "viftable2.csv")
```

**Model Validation**
```{r}
summary(testingSet1)
probsModel2 = as.numeric(unlist(predict(model2, testingSet1,type = "response")))

##accuracy check for testing data set
model2.pred = rep("No",218)
model2.pred[probsModel2>.5] ="Yes"
table(model2.pred, testingSet1$Pathogen2)
mean(model2.pred==testingSet1$Pathogen2)


##Area Under the Receiver Operator Characteristic Curve (AUROC)
chkroc2_test = pROC::roc(testingSet1$Pathogen2,probsModel2)
chkroc2_test

resiModel2 = as.numeric(unlist(residuals(model2)))

##correct classified for traininfset1
summary(trainingSet1)
probsModel2_train = as.numeric(unlist(predict(model2, trainingSet1,type = "response")))

##accuracy check for testing data set
model2.pred_train = rep("No",873)
model2.pred_train[probsModel2_train>.5] ="Yes"
table(model2.pred_train, trainingSet1$Pathogen2)
mean(model2.pred_train==trainingSet1$Pathogen2)


##Area Under the Receiver Operator Characteristic Curve (AUROC)
chkroc2_train = pROC::roc(trainingSet1$Pathogen2,probsModel2_train)
chkroc2_train
```


**Model diagnostics**
~~~~

~~~
```{r}

##Make predictions for total data sets
##making prediction with trainig data set
model2ProbsF = as.numeric(unlist(predict(model2, myDrilledWells,type = "response")))

resiModel2 = myDrilledWells$Pathogen1 - model2ProbsF
model2_diagnostics = data.frame(SortID,model2ProbsF,Pathogen1, Pathogen2,
resiModel2, WellDepth, CasingDepth,RatioDepthCasing2, Age, SepTankAF, ParcelSize,
xCoord, yCoord)

write.csv(model2_diagnostics, file = "model2_diagnostics.csv")
```

```{r}
d = read.csv("model2_diagnostics.csv")
summary(d)
ggplot(d, aes(xCoord, yCoord, colour =resiModel2 )) +
  viridis::scale_color_viridis()+
  geom_point(size = 3)
```
~~~